Bayesian Methods for Estimating Human Ancestry Using Whole Genome SNP Data

Claire Churchhouse

St. Hilda's College

Trinity 2012

Department of Statistics

University of Oxford

A dissertation submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

Supervisor: Jonathan Marchini

Bayesian Methods for Estimating Human Ancestry Using Whole Genome SNP Data

Claire Churchhouse, St. Hilda's College
Department of Statistics, University of Oxford
D.Phil. Thesis, Trinity 2012

Abstract

The past five years has seen the discovery of a wealth of genetics variants associated with an incredible range of diseases and traits that have been identified in genome-wide association studies (GWAS). These GWAS have typically been performed in individuals of European descent, prompting a call for such studies to be conducted over a more diverse range of populations. These include groups such as African Americans and Latinos as they are recognised as bearing a disproportionately large burden of disease in the U.S. population.

The variation in ancestry among such groups must be correctly accounted for in association studies to avoid spurious hits arising due to differences in ancestry between cases and controls. Such ancestral variation is not all problematic as it may also be exploited to uncover loci associated with disease in an approach known as admixture mapping, or to estimate recombination rates in admixed individuals. Many models have been proposed to infer genetic ancestry and they differ in their accuracy, the type of data they employ, their computational efficiency, and whether or not they can handle multi-way admixture. Despite the number of existing models, there is an unfulfilled requirement for a model that performs well even when the ancestral populations are closely related, is extendible to multi-way admixture scenarios, and can handle whole-genome data while remaining computationally efficient.

In this thesis we present a novel method of ancestry estimation named MULTIMIX that satisfies these criteria. The underlying model we propose uses a multivariate normal to approximate the distribution of a haplotype at a window of contiguous SNPs given the ancestral origin of that part of the genome. The observed allele types and the ancestry states that we aim to infer are incorporated in to a hidden Markov model to capture the correlations in ancestry that we expect to exist between neighbouring sites. We show via simulation studies that its performance on two-way and three-way admixture is competitive with state-of-the-art methods, and apply it to several real admixed samples of the International HapMap Project and the 1000 Genomes Project.

Acknowledgements

I would like to thank Jonathan Marchini for giving me the opportunity to work on this topic. It has been both challenging and rewarding and I have achieved far more under his supervision than I imagined I was capable of at the start of this process. I would like to thank my viva examiners, Simon Myers and Lachlan Coin, for their thoughts and discussion on this thesis. Many thanks to Garrett Hellenthal, Gil McVean and Chris Spencer for helpful comments at various stages of my work. I was fortunate enough to receive funding from the Department of Statistics throughout the course of my degree for which I am most grateful.

My officemates, past and present, have been helpful, supportive and tolerant of me on a day-to-day basis. Many thanks to A.M., E.F., V.I., M.F. and J.G. - I have been particularly lucky to have shared this experience with you. With so much of this work reliant on computing, I would also like to thank S. Hutchinson for always taking the time to help with my I.T. queries and for never considering any of my questions to be daft.

I have been lucky enough to have experienced more than the academic side of being an Oxford student. I would like to thank the teammates and friends that I've made through Oxford University Volleyball Club, Oxford University Women's Boat Club and St. Hilda's College Boat Club. The training, competition and laughter that I enjoyed with them was a healthy balance to my academic work and key in coping with the more trying days of my DPhil. Of all the wonderful things about life in Oxford, it is my friends that I'll miss the most. I would especially like to thank E.M., B.B. and B.P. with whom I've shared the fondest memories of my time here.

My interest in mathematics was always encouraged by my grandad, R.F. Churchhouse, whose passion and appreciation for numbers is highly contagious. I dedicate this thesis to my sister and my parents, who have shared in the whole experience with

me even though we live some three thousand miles apart. My Dad has always made my education a priority and without his support and encouragement I would not have had this opportunity and all the invaluable experiences that came along with it. My Mum has always been a refreshing voice that reminds me of the bigger picture when I lose perspective, and of belief in my abilities when I feel defeated. My big sister Kirsty, although we choose to travel down different paths, has always been there for me to help me along mine.

Contents

Chapter 1

Introduction and Survey of Existing Methods

1.1 The Global Extent of Human Genetic Variation

Admixture is defined as the interbreeding of two previously isolated populations (Bishop and Cannings, 2007). Prior to the admixture event there may have been geographical barriers between the two populations, such as a mountain range, desert or a large body of water, that have physically kept the populations apart for a long period of time. Alternatively, cultural and linguistic differences can prevent gene flow between two neighbouring populations (Barbujani and Sokal, 1991), maintaining their separation. The significance of the mixing from the point of view of genetics is that it brings together the genomes of individuals that have become, to some extent, genetically diverged from each other owing to their prolonged period of reproductive isolation. As a result, the admixed and subsequent generations will possess genomes that in some regions resemble those of one ancestral population and at other regions are more similar to those of the other.

Why is there genetic variation between populations? It is widely accepted that anatom-

ically modern humans evolved in Africa around 200,000 years ago (McDougall et al., 2005), and that approximately 50,000 - 100,000 years ago some migrated in waves from Africa and crossed in to Asia, Europe and eventually the Americas as they expanded across the globe according to the Out-of-Africa hypothesis of ancient human migration (Ramachandran et al., 2005). While the order in which the main geographic regions were occupied and the routes that were taken continues to be disputed, these migrants eventually inhabited the Old World and, by around 15,000 - 30,000 years ago, traveled across the Bering Straight to occupy the Americas. As some groups moved to new domains, a subset of the genetic variation that existed in the original population would be carried on to the dispersing group in what is known as the Founder Effect (Mayr, 1942). Following their expansion to different areas, these populations remained isolated from each other for tens of thousands of years. Over this time, the forces of mutation and genetic drift led to an increase in the variation between populations, as did selection which would have had effects dependent on the particular environmental and selective pressures acting on each population. Advances in seafaring, the establishment of global trade routes, colonization and the human slave trade have been some of the causes of the most recent admixture events bringing together people of different continents who, as a result, are to some extent genetically different to each other.

How much genetic variation is there between different populations? For over a century it has been known that there are differences between people in the types that they carry at certain genes. Following the discovery of the ABO blood group system, ABO gene frequencies were commonly used to classify populations after it was discovered that these blood types appeared at particular frequencies in different ethnic groups (Hirszfeld and Hirszfeld, 1919). Other genes whose types exhibit differences in the frequency at which they are found across the world include the hemoglobin gene whose rarer allele is responsible for sickle cell anaemia when in its homozygous form. This

is because individuals who are heterozygous for the gene benefit from an increased resistance to malaria which has been a selective pressure in specific parts of the world while not in others (Ringelhann et al., 1976).

While particular genes had been found to vary between populations, it was in the 1980s with the development of technologies that could read at the DNA level that a more complete picture of human genetic variation emerged. We now know that genetic variation comes in many flavours including microsatellites, biallelic single nucleotide polymorphisms (SNPs), insertions, deletions, inversions and copy number variants. We have learnt from studies such as the International HapMap Project (Altshuler et al., 2010) and 1000 Genomes Project (1000 Genomes Project Consortium, 2010) that most variants exist in all populations but often at different frequencies, and we have found very few that are *private*, that is, unique to a specific group. We now we have a much wider view of human genetic variation and can investigate global variation on a much finer scale - at the SNP level.

To quantify exactly how much of the variation that exists between individuals may be attributed to the fact that they belong to different populations we can calculate a measure of genetic distance known as F_{ST}. This statistic, first suggested by Sewall Wright (Wright, 1952), is the ratio of the between-population heterozygosity to the total heterozygosity of the two populations. It has been calculated pairwise between the European (CEU), West African (YRI), Chinese (CHB) and Japanese (JPT) populations of HapMap in Table 1.1 (Nelis et al., 2009). There are two things to note from these values: first, populations that are more closely located to each other tend to have lower F_{ST} than those that are further apart (although this is not always the case) and, secondly, all of these values show that the majority of genetic variation between individuals cannot be attributed to the fact that they belong to different populations. That is, when we speak of genetic variation between populations it really is a small extent of variation requiring sufficiently sophisticated statistical models and analytical techniques to learn

	YRI	CEU	CHB	JPT
YRI		0.153	0.190	0.192
CEU			0.110	0.111
CHB				0.007

Table 1.1: Calculation of the F_{ST} statistics between the European (CEU), West African (YRI), Chinese (CHB) and Japanese (JPT) populations of HapMap.

from it.

1.2 Genomes of Admixed Individuals

The chromosomes of admixed individuals, who have recent ancestry derived from two or more genetically diverged populations, may be thought of as being composed of segments of distinct ancestry. If the admixture involves two distinct populations then it is referred to as *two-way* admixture, and if a third population is also involved in the mixing it is *three-way* admixture. When speaking of *multi-way* admixture, we are referring to that which involves three or more populations. The sites of switches in ancestry are a result of recombination events that have brought together chromosomes of different population origins in meiotic crossover of the admixed individual's ancestors. The length of these segments will depend on the number of generations since the admixture occurred, with shorter segments indicating older admixture events, the relative number of individuals in each population and whether or not gene flow continued over a period of time. The distribution of these ancestral chunks across the genome will vary among individuals, even for those who identify as sharing a common ancestral background. This genetic ancestry is not directly observable and there are many motivations for developing methods to accurately infer it.

To see how these chunks arise consider that for a certain individual, who will be central to the following explanation, their maternal grandparents came from two different populations - the maternal grandmother belonging to population 1 and the ma-

ternal grandfather to population 2. All alleles originating from population 1 will be coloured red and those originating from population 2 are coloured blue in Figure 1.1. The individual's grandparents belong to the generation in which the admixture event occurred and when they reproduce each one of them will donate a chromosome copy that is entirely of genetic material derived from their respective populations. As a consequence, their offspring (the individual's mother) has inherited one copy from population 1 and the other from population 2. In this first generation since mixing, all of the chromosomes are entirely from one population or the other. It is only during the meioses within the second generation (that to which the parents belong) that chromosomes of different ancestral origins are fused together during recombination, resulting in copies that are a hybrid of both ancestral sources. The black line that traces initially through the mother's red copy and then makes a switch to her blue copy, depicts the transmitted DNA that is then donated to the individual as the offspring. The site at which the black line jumps from the red to the blue copy is a recombination event which in this case leads to an ancestral switch in child's copy. The location of these recombination events and how many occur in each meiosis is random, with some regions of the genome, known as hotspots, being particularly active and others rarely featuring such events (Myers et al., 2005, 2008).

If we imagine that the individual's father shared a similar genetic history to that of their mother, inheriting one copy from each population, then the paternal copy that the individual possess will also be an admixed copy. Once again, the cross-over path will jump between the two copies and in this case each of the two jumps results in a switch in ancestry in the transmitted copy that is passed on to the individual's child, although this need not be the case. Note that if the path jumped from say a red site on one copy to another red site on the other, then a recombination event would have occurred but it would not have left any evidence of it having taken place. That is, the location of the resulting switches in ancestry are a subset of the total set of sites at

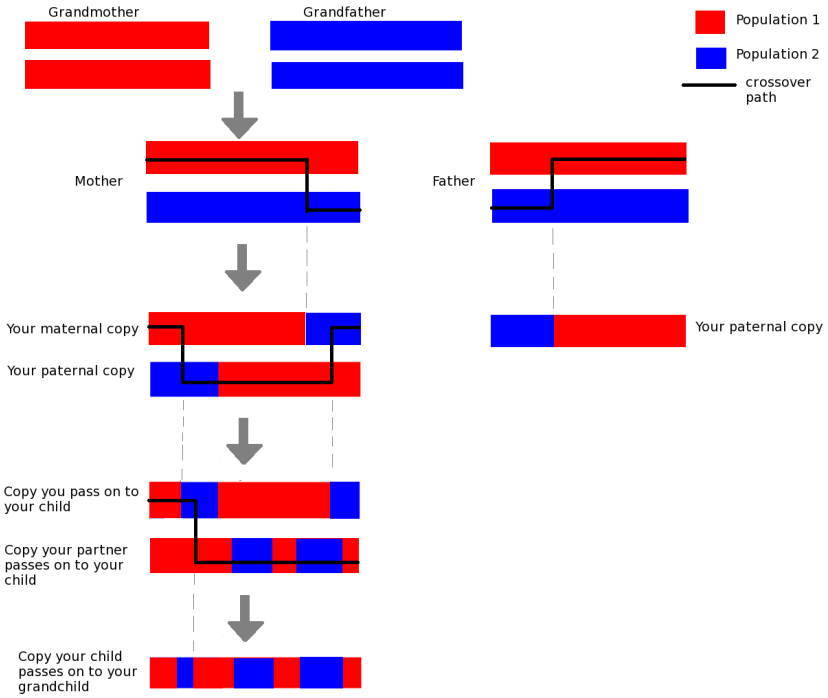

Figure 1.1: A schematic illustration of how chunks of ancestry arise as a result of admixture between two distinct populations. With increasing number of generations since the admixture occurred, there are more recombination events that can lead to the joining together of haplotypes of different ancestral origins resulting in shorter chunks and more switching. The colours in this illustration are the unobservable states pertaining to the population of origin of each allele that ancestry estimation techniques aim to infer.

which recombination events have happened.

We can see how over generations since the mixing, there are more and more recombination events along admixed chromosomes generating increasingly small ancestral chunks in the genomes of subsequent generations. It is this pattern of alternating coloured blocks that ancestry estimation in admixed individuals aims to uncover, effectively painting each chromosome according to its population of origin at each site.

1.3 Data on Worldwide Genetic Variation

Thanks to collaborative efforts, there have been several large-scale projects to catalogue human genetic variation in different populations around the world, for which the data have been made freely available.

- **HapMap3**: The International HapMap Consortium is one such project involving scientists from Canada, China, Japan, Nigeria, the United Kingdom and the United States. In its third and most recent phase, HapMap3, it has collected data consisting of 1,184 reference individuals from 11 global populations : African ancestry in Southwest USA (ASW); Utah residents with Northern and Western European ancestry from the CEPH collection (CEU); Han Chinese in Beijing, China (CHB); Chinese in Metropolitan Denver, Colorado (CHD); Gujarati Indians in Houston, Texas (GIH); Japanese in Tokyo, Japan (JPT); Luhya in Webuye, Kenya (LWK); Mexican ancestry in Los Angeles, California (MXL); Maasai in Kinyawa, Kenya (MKK); Toscani in Italy (TSI); Yoruba in Ibadan, Nigeria (YRI) (Altshuler et al., 2010). These samples were genotyped on two different platforms (the Affymetrix Human SNP array 6.0 and the Illumina Human 1M-single beadchip) giving a merged set at over 1.6 million SNPs across the genome. Some regions were then selected for resequencing in about half of the samples, leading to the detection of rarer SNPs and a more accurate picture of the allele frequency spec-

trum for each population. The outcome of this project is a large high-quality set of haplotypes that captures patterns of variation in 4 African, 2 European and 2 Asian populations as well as 2 admixed groups - the Mexicans and the African Americans (Altshuler et al., 2010).

- **The Human Genome Diversity Project (HGDP)**: This data set consists of samples from 1,050 individuals from a wider variety of 52 populations including Native American, Middle Eastern and Oceanic groups (Cann et al., 2002). Compared to HapMap, the HGDP panel covers a wider spread of populations and includes individuals from Oceania, the Americas, the Middle East and Central and South Asia. These samples were genotyped at approximately 650,000 common SNPs and as with HapMap their data were made freely available making it an invaluable database for studies of human evolution and genetic anthropology.

- **The 1000 Genomes Project**: The most recent large-scale study of worldwide patterns of variation is the 1000 Genomes Project. It is the first project to provide a resource of human genetic variation through next-generation sequencing technologies, aiming to find most genetic variants that appear at a frequency of at least 1% in each of the populations investigated (1000 Genomes Project Consortium, 2010). The samples included in the project come from populations across Europe, South and East Asia, West Africa and the Americas. In addition to featuring African Americans and Mexicans, they also include admixed groups such as Puerto Ricans, Colombians and African Caribbeans. In all of these studies, the people from whom samples were taken have been carefully selected to be representative of the population to which they belong by requiring, for example, that all four of their grandparents were also from that particular region. With such a broad resource of global human genetic variation available, we are now able to make comparisons between the genome of an individual of unknown ancestry to

these panels of samples through statistical techniques and draw inference about their ancestral background.

1.4 mtDNA and Y-chromosome Haplogroups

The analysis of mitochondrial DNA (mtDNA) to learn about the maternal lineage of an individual, and the study of haplogroups along the non-recombining portion of the Y-chromosome to ascertain the paternal history of the subject were popular approaches early on in studies of global origins in the 1980s and 1990s (Cann et al., 1987). Since neither of these portions of DNA recombines, they remain unchanged throughout many generations with new variants arising only very rarely due to chance mutations making them suitable for identifying stable maternal or paternal lineages that can be traced back in time over thousands of years (Hammer et al., 2002). People may be classified as belonging to a particular haplogroup if they share a mutation on a certain haplotypic background, indicating a common ancestor dating back thousands of years.

Currently, the Y Chromosome Consortium identifies a tree of over 300 distinct haplogroups constructed from over 600 binary markers that fall in to 20 major clades of ancestry (Karafet et al., 2008). These clades are defined by a certain number of mutations and the haplogroups that they encompass are typically found in or restricted to particular geographic regions. For instance, Clade A which is the most basal clade in the Y chromosome tree of Karafet et al. has been found almost exclusively within the African continent with its haplogroups occurring most frequently in Ethiopian, Sudanese and Khoisan individuals. The lineages of Clade D, which is defined by two mutations, are most commonly found in Central Asia in Japan but are also present at low frequencies in Southeast Asia, while those of Clade G are distributed less widely, being present mostly in the Middle East, the Mediterranean, and the Caucasus Mountains.

While mtDNA and haplogroup studies are informative in their own right, each of these approaches only produces a partial view of the genetic ancestry of an individual as they consider, in isolation, the mitochondrial chromosome and Y chromosomes respectively. It is only females that pass on mtDNA. Y chromosomes, being possessed solely by males, are exclusively passed down a male lineage. They do not tell us about the origins of the alleles that comprise the 22 pairs of autosomes. In admixed individuals, where the autosomal ancestry will vary in a mosaic-like pattern, mtDNA and haplogroup analyses are particularly insufficient as they do not elucidate this ancestral switching. With the surge of genome-wide SNP data becoming available for populations sampled from across the globe, interest has grown in the development of statistical methods to utilize this richer data for ancestry estimation, allowing us to consider the ancestral history of the whole genome simultaneously.

1.5 Ancestry Informative Markers

The large-scale studies described in section 1.3 have provided a platform upon which to compare the allele frequencies between different populations at a vast set of SNPs across the genome. With such resources available, one can pick out which SNPs are putatively informative for ancestry in admixed individuals by identifying those that show marked differences in their allele frequencies between the ancestral populations. Such SNPs are referred to as ancestry informative markers (AIMs).

A study by Price et al. (Price et al., 2007) sought to construct an *admixture map* of AIMs that are informative of ancestry in Latinos by screening multiple datasets for SNPs that give some indication of the ancestral origin of their alleles. Specifically, their goal was to ascertain an admixture map of SNPs that were telling of whether their alleles were of Native American ancestry as opposed to European or African, these being the common sources of the genetic origins of most Latinos (Salzano and Bortolini,

2002). SNPs were selected to be part of the map based upon their expected Shannon information content (SIC) between European and Native American populations. Markers were iteratively added to the map if, when taking in to account the information already captured by the map, they were the most informative according to their SIC prediction and were no less than 0.3cM from previously selected sites. From a collated database of a few million markers, they found 1,649 loci that fit this purpose and were validated in a separate Latino dataset.

The motivation behind the construction of this admixture map was for its use in admixture mapping studies, which we will discuss in the following section. Admixture mapping requires that you know which populations are ancestral to the admixed individuals being studied. Futhermore, AIMs must be identified separately for every new combination of populations contributing to the admixture meaning that they are very application-specific and it becomes increasingly complex to find AIMs when considering multiple populations rather than just two. Admixture maps are usually composed of no more than a few thousand loci genome-wide that satisfy the requirement of exhibiting substantially different allele frequencies in the ancestral populations, meaning that the inferred local ancestry of the admixed samples is incomplete. In order to uncover a local and whole-genome picture of genetic ancestry at a set of loci dense enough to infer where the switches in ancestry occur, another approach is required.

1.6 The Importance of Ancestry Estimation

Ancestry inference of an individual's genome may be carried out at two levels - globally or locally. *Global* ancestry refers to the overall proportion of an individual's genome derived from a particular ancestral population while the more descriptive *local* ancestry specifies the population of origin of each allele for every locus being considered. In case-control association studies and admixture mapping, knowledge of the local

ancestry of the subjects is required.

In case-control association studies it is important to correct for population structure that may otherwise lead to spurious associations between a genotype and disease status (Marchini et al., 2004). In this context, the estimated proportions of an individual's genome inherited from each population may be included as covariates in a linear regression model of genotype on disease status where self-reported ancestry is not always reliable as it may differ considerably from the true genetic ancestry. With the increasing number of genome-wide association studies (GWAS) being carried out in African American and Latino populations (Rosenberg et al., 2010), accurate ancestry estimation techniques are essential to correctly account for the genotypic variation attributable to ancestry.

Admixture mapping is a technique that seeks to gain from the variation in ancestry among admixed disease cases. In admixture mapping, we search for a region in the genomes of admixed cases in which there is an unusually high proportion of ancestry from the population with the higher incidence of the disease. Under the assumption that differences in disease prevalence between populations are, to some extent, due to the disease-causing variant being at different frequencies in the populations, scanning the genome in this way can identify variants associated with the disease. This approach has been successful in locating genes associated with diseases such as type 2 diabetes (Elbein et al., 2009; Kao et al., 2008), breast cancer (Fejerman et al., 2009), prostate cancer (Bock et al., 2009; Freedman et al., 2006) and quantitative traits including hypertension (Zhu and Cooper, 2007; Zhu et al., 2005) and obesity (Cheng et al., 2009, 2010).

By uncovering the genetic ancestry of a population, we can learn of its demographic history and ask questions about how is it related to other groups, whether there were any major events such as bottlenecks and if it is admixed then what ancestral populations were involved in the admixture event and how far back in time did they mix? J.E. Pool and R. Nielsen (Pool and Nielsen, 2009) proposed a model to describe the distri-

bution of the length of ancestry chunks inherited from one population via migration at a constant rate in to another. An extension of this model allows for the migration rate to have changed some number of generations ago. By employing a method of local ancestry estimation to deduce the distribution of ancestry chunks within a sample, this information can then be used to estimate the migration rates and the time since there was a change in these rates, if at all, allowing different hypotheses of demographic history to be tested. Another more recent example of local ancestry estimation being used to learn of a population's demographic history is a genotyping and resequencing study of 22 Mexican Americans aimed at investigating the Native American component of the samples (Wall et al., 2011). By first inferring among the samples which parts of their genomes each had inherited from Native American ancestors, they then exclusively studied these Native American components using computational approaches to deduce that a population bottleneck occurred around 12,500 years ago, a time that is consistent with archaeological estimates of the time of human migration to the Americas.

We can also look for signatures of recent selection in admixed populations by asking whether or not there are regions of the admixed genomes at which there is a significant excess or deficit in ancestry from a particular population, as compared with the genome-wide average. Such studies were carried out in African Americans (Bryc et al., 2010a) and Puerto Ricans (Tang et al., 2007) although neither have returned evidence of selection that maintains significance after correcting for genome-wide testing or could not alternatively be explained by inadequate modeling of long-range LD (Price et al., 2008).

Recently, inference of genetic ancestry in admixed individuals has been leveraged to deduce recombination maps in populations such as African Americans (Hinch et al., 2011; Wegmann et al., 2011) and African Caribbeans (Wegmann et al., 2011). Recombination rates have typically been calculated in European populations where detailed ge-

nealogical information is available from large accurate pedigrees such as the Icelandic pedigree that produced the deCODE map (Kong et al., 2010). Taking a novel approach, this work estimated genome-wide recombination rates by inferring the number and location of switches in ancestry, and hence recombination events, in a large sample of individuals. Studying admixed samples was efficient as it provided about three times as many observable recombination events as compared to the number of informative recombinations inferred through the pedigree (Wegmann et al., 2011). In both of these studies, HAPMIX (Price et al., 2009) was the method of choice to infer ancestry. The construction of these admixed recombination maps has shed light on how recombination rates differ between populations. Hinch et al. found that while 2,500 hotspots were unique to regions of African ancestry, no hotspots were discovered to be unique to European ancestry. It follows that it is not only the larger effective population size of Africans but this higher recombination activity that is a cause of LD extending over shorter distances in Africans as compared to in Europeans.

Aside from the applications of ancestry estimation to medicine, population genetics and demographic history, there is also a personal interest in genetic ancestry on which some direct-to-consumer (DTC) genomics companies aim to capitalize. These companies offer ancestry estimation as a service to their clients, with California-based 23andMe for example providing a colour-coded illustration or "chromosome painting" (Fig.1.2) of the genome allowing people to see which parts they have inherited from ancestors of various populations. A notable limitation of the analysis provided by 23andMe is it only makes use of three populations as possible ancestral sources - a European, an African and an Asian group. While this may be sufficient for studying the genomes of African-American clients, it is less than ideal in its application to people who may have Native American or Latino ancestries as there are more representative populations available in these cases for which genetic data has also been collected in studies such as the HGDP.

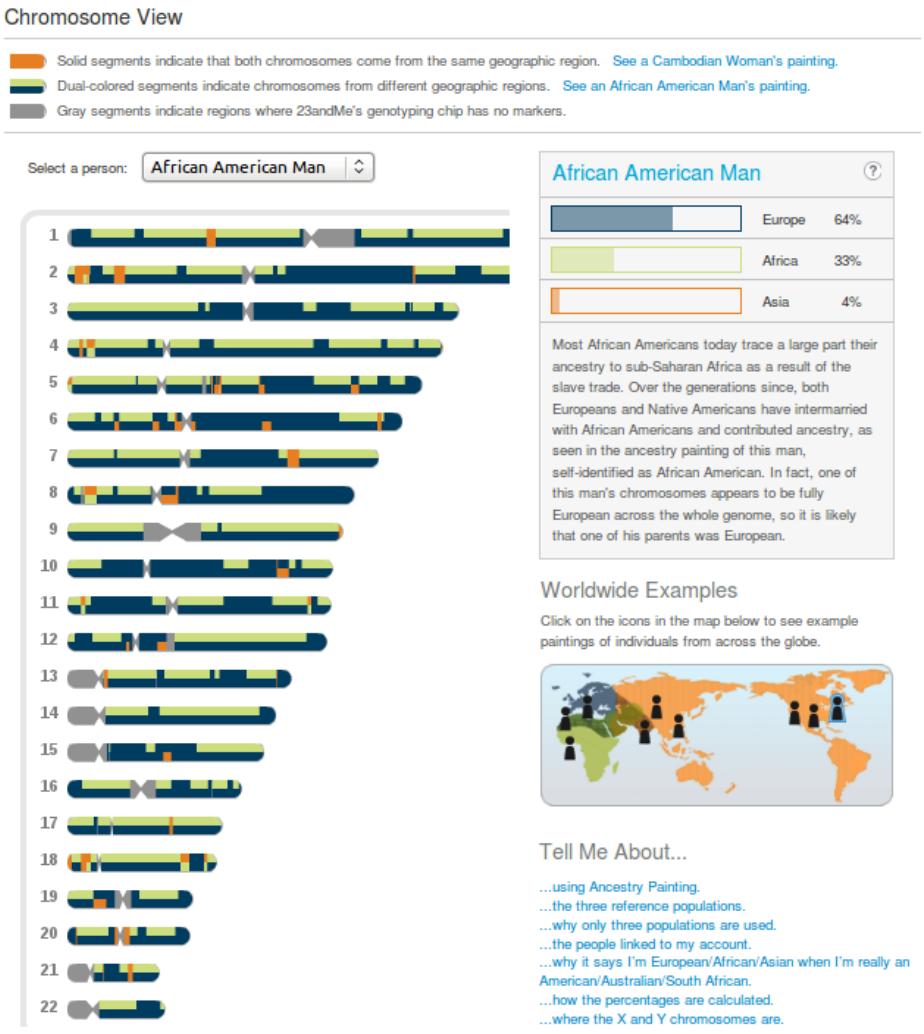

Figure 1.2: An example of the chromosome painting of an African American as presented by the direct-to-consumer genetic testing service 23andMe. Each part of the autosomes is painted a colour corresponding to its inferred population of origin.

1.7 Existing Methods

There is a wealth of ancestry estimation methods in the literature and here we describe some of the most notable. They employ a range of statistical techniques and differ in the exact motivation for the ancestry estimation as well as in the scenario to which they may be applied. We have chosen to include the following approaches in our survey of methods to demonstrate how the field has developed upon existing techniques, and to present those that have been successfully adopted for use in numerous studies. Furthermore, all of the novel methods discussed here are applicable to the analysis of SNP data making them comparable to the methods that we propose in this thesis in Chapter 2. We progress through these ancestry estimation tools in chronological order, giving an explanation of the underlying models and techniques of parameter estimation. A comparison of the key features of these methods is summarized in Table 1.2.

1.7.1 STRUCTURE

One of the earliest methods, Structure (Pritchard et al., 2000) employs a model-based Bayesian clustering technique to infer population structure and assign individuals to populations probabilistically. As the approach is Bayesian, it nicely incorporates the uncertainty of the parameter estimates in to the inference. Structure may be applied when we can assume there are K populations, each with its own set of characteristic allele frequencies at a set of unlinked markers. It attempts to assign individuals to a population on the basis of their genotypes while, at the same time, estimating the allele frequencies in the different populations based upon this membership. Hardy-Weinberg equilibrium is assumed within each population, thus the likelihood of the observed genotypes given the population of ancestry is simply the product of the corresponding allele frequencies in that population.

Given the genotypes of the individuals, X, knowledge about the ancestry Z and the

Method	Applicable to > 2 populations?	Does it model background LD?	Requires phased ancestral data?	Notes
STRUCTURE linkage model (Pritchard et al., 2000; Falush et al., 2003)	Yes	No	No	HMM on the ancestry of alleles at unlinked SNPs with MCMC sampling to estimate local ancestry.
ANCESTRYMAP (Patterson et al., 2004)	No	No	No	HMM designed for admixture mapping of disease genes.
FRAPPE (Tang et al., 2005)	Yes	No	No	Uses the original STRUCTURE model with EM implementation to estimate global ancestry proportions.
SABER (Tang et al., 2006)	Yes	1st-order Markov HMM	Yes	Accounts for background LD to some extent by considering pairwise allele frequencies when no change of ancestry is inferred.
LAMP (Sankararaman et al., 2008b)	Yes	No	No	A clustering algorithm on windows of SNPs that analyses several admixed individuals at once and does not require the ancestral genotypes.
HAPAA (Sundquist et al., 2008)	Yes	Yes	Yes	A hierarchical HMM that assumes pre-phasing of the sample data and includes a heuristic algorithm to identify sites of phase-switch errors.
SWITCH and SWITCH-MHMM (Sankararaman et al., 2008a)	Yes	SWITCH-MHMM is a 1st-order Markov HMM	Yes	Explicitly models sites of recombination events, upon which the emission probabilities are conditional rather than upon the ancestry states themselves.
WINPOP (Pasaniuc et al., 2009)	Yes	No	No	An improvement of lamp with an adaptive window size that varies with genomic location.
HAPMIX (Price et al., 2009)	No	Yes	Yes	An extension of the Li and Stephens model of recombination to a two-way admixture scenario.
ADMIXTURE (Alexander et al., 2009)	Yes	No	No	A more computationally efficient implementation of the STRUCTURE model that estimates global ancestry proportions.
RFmix (Bryc et al., 2010a)	Yes	No	No	A HMM approach where emission probabilities are derived from the PCA loadings and the individual's genotype across windows of SNPs. Transitions between ancestry states are assumed to be Poisson distributed.
LAMP-LD and LAMP-HAP (Baran et al.)	Yes	Yes	Yes	An HMM approach that captures the haplotype structure of the ancestral populations via an approximation of the Li and Stephens model that reduces the number of possible states, making it faster to implement than HAPMIX.
MULTIMIX	**Yes**	**Yes**	**No**	A Multivariate Normal model on haplotype probabilities given ancestry and a HMM on how ancestry changes along a chromosome.

Table 1.2: A comparison of ancestry estimation techniques. MULTIMIX is the only method that models background LD, is applicable to more than two populations and does not require that the ancestral haplotypes are phased.

population allele frequencies P is given by:

$$Pr(Z, P|X) \propto Pr(X|Z, P)Pr(Z)Pr(P)$$

i.e. the posterior probability is proportional to the product of the likelihood and the prior distributions. An approximate sample from $Pr(Z, P|X)$ is obtained via MCMC by sampling the pairs $(Z^{(1)}, P^{(1)}), ..., (Z^{(M)}, P^{(M)})$ over M iterations.

Two scenarios are considered - the first in which each individual is assumed to originate from one of the K populations, and the second in which individuals can have partial ancestry from more than one population (i.e. individuals may be admixed). The distribution of the frequency of the alleles at locus l in population k is modeled as a Dirichlet distribution, $p_{kl.} \sim D(\lambda_1, \cdots, \lambda_{J_l})$ where there are J_l different allele types at locus l.

In the first scenario, the model without admixture, Gibbs sampling is applied as follows:

Initialize $Z^{(0)}$ by randomly drawing the ancestral population of each individual from $Pr(Z^{(0)} = k) = \frac{1}{K}$

Step 1 - Sample $P^{(m)}$ from $Pr(P|X, Z^{(m-1)})$ where

$$p_{kl.}|X, Z \sim Dirichlet(\lambda_1 + n_{kl1}, ..., \lambda_{J_l} + n_{klJ_l}) \tag{1.1}$$

and n_{klj} is the number of copies of allele j at locus l that have been observed in the individuals currently assigned by Z to population k.

Step 2 - Sample $Z^{(m)}$ from $Pr(Z|X, P^{(m)})$ where

$$Pr(z^{(i)} = k|X, P) = \frac{Pr(x^{(i)}|P, z^{(i)} = k)}{\sum_{k'=1}^{K} Pr(x^{(i)}|P, z^{(i)} = k')} \tag{1.2}$$

and

$$Pr(x^{(i)}|P, z^{(i)} = k) = \prod_{l=1}^{L} p_{klx^{(i,1)}} \cdot p_{klx^{(i,2)}} \tag{1.3}$$

the product of the frequencies of alleles $x^{(i,1)}$ and $x^{(i,2)}$ (of individual i) at locus l in population k. Iterating the sampling of P in step 1 and Z in step 2 produces pairs of (Z, P) values from the target distribution, $Pr(Z, P|X)$.

In the second scenario, the model is extended to account for admixed individuals, and a new variable Q is used to denote the admixture proportions of each individual. The MCMC sampling involves the following steps:

Step 1 - Sample $P^{(m)}$ and $Q^{(m)}$ from $Pr(P, Q|X, Z^{(m-1)})$. P is sampled as before, Q is sampled from

$$q^{(i)}|X, Z \sim Dirichlet(\alpha + m_1^{(i)}, ..., \alpha + m_K^{(i)}) \tag{1.4}$$

where $m_k^{(i)}$ is the number of allele copies the ith individual that descended from population k according to the current Z.

Step 2 - Sample $Z^{(m)}$ from $Pr(Z|X, P^{(m)}, Q^{(m)})$.

$$Pr(z_l^{(i,a)} = k|X, P) = \frac{q_k^{(i)} \cdot Pr(x_l^{(i,a)}|P, z_l^{(i,a)} = k)}{\sum_{k'=1}^{K} Pr(x_l^{(i,a)}|P, z_l^{(i,a)} = k') \cdot q_{k'}^{(i)}} \tag{1.5}$$

where $z_l^{(i,a)}$ is the population of origin of the allele $x_l^{(i,a)}$, and

$$Pr(x_l^{(i,a)}|P, z_l^{(i,a)} = k) = p_{klx_l}^{(i,a)} \tag{1.6}$$

Step 3 - Update α where $q^{(i)} \sim Dirichlet(\alpha, ..., \alpha)$ using a Metropolis-Hastings update, i.e. simulate a proposal α' from $N(\alpha, \sigma_\alpha^2)$ and reject α' if $\alpha' \leq 0$, otherwise accept with MH probability.

1.7.2 STRUCTURE Linkage Model

The widely-used linkage model (Falush et al., 2003) accounts for correlations between linked loci that are present in admixed populations. The original Structure model assumes that the ancestry states, the denoted Z, are independent along a chromosome, so it ignores the correlations in ancestry that one would expect to see for loci that are close enough to each other. The second model, Structure 2, infers the ancestral population of chromosomal regions and allows for the detection of even older admixture events.

In the admixture model (the second scenario presented above) of Structure 1, each allele copy is independently derived from one of the K populations. Alternatively, in the linkage model, it is blocks of chromosomes that are inherited from one of the ancestral populations. They assume that the breakpoints (i.e. loci at which the ancestral population changes) between contiguous chunks may be modeled as a Poisson process of rate r per unit genetic distance. As this rate r becomes increasingly large, all loci become independent and the model resembles that of Structure 1. The population of origin of each chunk is randomly drawn according to the overall admixture proportions, q.

Under this model, the ancestry states Z form a hidden Markov model; the population of origin at each loci is unobservable and hence is the hidden state in the Markov chain. They assume that a sample is drawn from a diploid population which has, at some point in its history t generations back in time (of which r may be considered an estimate), experienced a single admixture event. After this event, a number of generations of random mating (within the admixed population itself) have occurred resulting in a population of individuals, some of which constitute our sample. In the generation in which the admixture occurred, the individuals inherit their DNA "intact" as there is no recombination. In the following generation however, cross-over events in a meiosis

will result in boundaries delineating these chunks of ancestry where these breakpoints occur as a Poisson process of rate 1 per morgan. As this process is repeated at every subsequent generation, the superposition of these Poisson processes results in the breakpoints along the chromosomes of the current generation forming a Poisson process of rate t per morgan.

Another new feature of the second Structure model is that its prior model accounts for correlations between the allele frequencies that occur in closely related populations, where the allele frequencies may be very similar. A Dirichlet prior whose parameters are a function of the allele frequencies in the hypothetical ancestral populations (from which the K populations in the sample have experienced genetic drift) and the rate of drift away from these ancestral allele frequencies is used.

As in Structure 1, MCMC sampling is used although this time the target distribution is $Pr(P, Z, r, Q|X, K)$. The sampling scheme may be summarized in the following steps:

Step 1 - Sample $Pr(Z|P, r, Q, X)$ by the forward-backward algorithm.

Step 2 - Sample $Pr(P|Z, r, Q, X) = Pr(P|Z, X) \sim$ Beta distribution.

Step 3 - Update r with a Metropolis-Hastings update.

Step 4 - Update Q with a Metropolis-Hastings update.

1.7.3 ANCESTRYMAP

ANCESTRYMAP is a method that was designed to estimate local ancestry in a sample of individuals who are two-way admixed, with the aim of performing admixture mapping to scan for disease genes. For the admixture mapping to have sufficient power, it requires that the loci used in the model show considerable differences in allele frequencies between the two populations of the admixture. A panel of 2,154 such markers across the genome was constructed for use in admixture mapping in African Ameri-

cans where the SNPs included showed an average difference in allele frequency of 57% between European Americans and West Africans (Smith et al., 2004).

The underlying model is the same as the HMM of Structure 2, but uses adaptive-rejection sampling to speed up computation with MCMC sampling accounting for the uncertainty of the unknown variables : global ancestry proportions, number of chromosomal exchanges per morgan between ancestral segments of the genome since the mixing event, and the allele frequencies in each population. These parameters are initialized at reasonable values, depending on what is already known about the admixed samples being analysed, in order to reduce the burn-in period of the MCMC sampling. For example, the allele frequencies may be initialized at the values estimated from the parental populations, say European Americans and West Africans in the case of African American samples. The proportion of ancestry each individual possess from the two ancestral populations may be initialized at its maximum-likelihood estimate arrived at by naively treating all SNPs as unlinked and ignoring the expected correlations in ancestry among nearby sites.

One of two possible statistics, the *locus-genome* statistic or the *case-control* statistic, is calculated and averaged over all iterations after burn-in to test each locus for disease association while accounting for uncertainty in the model parameters. The first of these statistics is applied when the admixture mapping involves only cases of the disease, as it is used to test whether any region among these cases exhibits a significantly high proportion of ancestry from one of the ancestral populations as compared with the average proportion seen across the admixed genomes. The second statistic is used when analysing both cases and controls, comparing the ancestry estimates between these two groups at every locus. It is used to test for a significant deviation from the genome-wide average of ancestry of one population that is evident in the cases but absent in the controls.

1.7.4 FRAPPE

This work is a frequentist take on the Bayesian approach of Structure. It requires that markers are unlinked and is designed to estimate the global ancestral proportions of an individual. As with Structure, ancestral allele frequencies are estimated using data from both the ancestral and the admixed samples.

An EM algorithm is designed to update estimates of allele frequencies P and global ancestry, Q with the local ancestry states being the unobservable variable Z. The likelihood function of the parameters P and Q given the genotype data G and hidden states Z is the product of the corresponding allele frequencies and global proportions over all individuals, for every locus. That is

$$logL(G, Z|P, Q) = \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{a=1}^{2} \sum_{l=1}^{L_m} \sum_{k=1}^{K} \mathbb{1}[G_{ima} = l, Z_{ima} = k] \cdot log(p_{mlk} \cdot q_{ik})$$

Here I is the number of individuals (both ancestral and admixed), M is the number of loci, L_m is the number of alleles at locus m, K is the number of populations, and the product over index a accounts for each of the two alleles at a marker.

At any iteration n, the expectation of local ancestry for allele a of individual i at locus m for population k is denoted

$$\mathbb{E}_{imak}^{(n)} = \mathbb{E}(\mathbb{1}[Z_{ima} = k]|P^{(n)}, Q^{(n)}, G)$$

To initialise the algorithm, the values of $\mathbb{E}_{imak}^{(0)}$ for the ancestral individuals are set to 1 for k equal to the population to which they belong, and 0 otherwise. For the admixed individuals, a random assignment of $\mathbb{E}^{(0)}$ that satisfies $\sum_{k=1}^{K} q_{ik} = 1$ sets the initial expectations.

Maximization step - The maximization step computes the MLE of P and Q conditional on the current expectation of Z. The update of allele frequencies at iteration n

is

$$p_{mlk}^{(n)} = \frac{\sum_{i=1}^{I}\sum_{a=1}^{2} \mathbb{1}[G_{ima} = l]\mathbb{E}_{imak}^{(n-1)}}{\sum_{i=1}^{I}\sum_{a=1}^{2} \mathbb{E}_{imak}^{(n-1)}}$$

Ancestral proportions are updated by

$$q_{ik}^{(n)} = \frac{\sum_{m=1}^{M}\sum_{a=1}^{2} \mathbb{E}_{imak}^{(n-1)}}{2M}$$

Expectation step - For the current estimates of P and Q, Bayes rules gives the expectation of the missing variables:

$$\mathbb{E}_{imak}^{(n)} = \frac{P(G_{ima}|Z_{ima} = k) \cdot P(Z_{ima} = k)}{P(G_{ima})} = \frac{p_{mlk}^{(n)} q_{ik}^{(n)}}{\sum_{k'=1}^{K} p_{mlk'}^{(n)} q_{ik'}^{(n)}}$$

The expectation and maximization steps are repeated until the absolute change in parameter estimates is less than some threshold value.

1.7.5 SABER

SABER (Tang et al., 2006) uses a first-order Markov-hidden Markov model (MHMM) to account for the background linkage disequilibrium that HMM methods ignore. The emission probabilities at each locus will be the allele frequency in the relevant population (as in most HMM admixture models) if there is a change in ancestry state at the locus, however if there is no such change then the emission probability will be the two-marker haplotype frequency instead. That is, if $X_i \in \{0, 1\}$ is the observed haplotype

at locus i and Z_i is its population of ancestry then

$$P(X_i|Z_1,\ldots,Z_i,X_1,\ldots,X_{i-1}) = \begin{cases} P(X_i|Z_i,X_{i-1}) & \text{if } Z_i = Z_{i-1} \\ P(X_i|Z_i) & \text{otherwise} \end{cases}$$

The transition probabilities in the MHMM are based upon the expected length of ancestral blocks, allowing the model to account for different admixing times for 3 of more ancestral populations. They consider τ_k, where k is the population number, to be the inverse of the expected length of ancestral blocks that are derived from population k. The "instantaneous" transition rate from ancestral state i to state j (Q_{ij}) is defined such that, given the current state, the length until the next point of recombination (which may or may not result in a change in ancestry) is exponentially distributed which an expectation that is inversely proportional to the number of meioses since admixing. Thus, the more generations since the admixture event of population k occurred, the shorter the expected length of ancestry blocks for population k. The authors also observe that the probability of switching in to a particular state should be inversely proportional to the expected time that the Markov chain remains in that state. The transition matrix is computed from this "instantaneous" transition rate matrix to be $e^{-(d*Q)}$ where d is the distance in Morgans between two markers.

A forward algorithm is used to compute the log-likelihood of the set of parameters given the observed data, and has a formulation dependent on whether the data is phased or not, as does the backward algorithm. The algorithm used for sampling ancestral states from the posterior distribution is similar to the backwards Gibbs sampling of Structure 2.

1.7.6 LAMP

LAMP (Sankararaman et al., 2008b) is a non-model based approach that slides a window of optimal length, l, along a stretch of unlinked SNPs and at every position of the window and uses a clustering algorithm to estimate the ancestral population within the window. Every SNP will be covered by many windows, and a majority vote over the ancestral populations estimated by all windows that contain the SNP establishes the most likely population of ancestry. The optimal window length is estimated in the model such that it is short enough so that most individuals will have no breakpoints in ancestry within the window, but long enough to encompass enough information so that it may correctly cluster the individuals to their ancestral population(s) via the clustering algorithm.

The method tries to minimize the errors in ancestry estimated at each SNP or locus rather than trying to estimate the exact location of breakpoints in ancestry, as in Structure and SABER. It does not require the ancestral genotypes to be known, although there is an alternative version of the method, LAMP-ANC, which utilizes the pure ancestral genotypes when they are available. The modeling assumptions include assuming that K ancestral populations have been mixing for g generations, of recent admixture, and that they are known. These populations are allowed to have mixed at different times however, so g is the upper bound on the number of generations since admixture, and random mating with the whole pool of populations at every generation is also assumed. Unlike in Structure and SABER, the admixed proportions q are also considered to be known. The rate of recombination in any one meiosis is assumed to be the same at all SNPs, $r_j = r$. The transmission of a chromosome from a parent to their child is modeled as tracing along a chromosome from one end to the other where the cross-overs between chromosomes occur as a Poisson process of rate r.

Based upon the parameters g, q and r, they calculate an optimal window length

l and use a clustering algorithm of Iterated Conditional Modes (ICM) to find the best classification of ancestral population for each individual in terms of the likelihood. The ICM updates are similar to those of the EM algorithm, except in the expectation step, where rather than finding the expected classification θ (given the allele frequencies and the individual's genotype) they seek the maximum a posteriori (MAP) estimate of θ, as this avoids fractional class membership that may be returned by a standard E-step. In the M-step, the maximum-likelihood estimates of the allele frequencies in each population, p_k, are sought. They consider two scenarios, characterized by how many ancestral populations there are and whether or not the p_k are known, each scenario calling for its own initialization of the p_k and the classification function θ.

1.7.7 HAPAA

The HAPAA method (Sundquist et al., 2008) stands for the **H**MM-based **A**nalysis of **P**olymorphisms in **A**dmixed **A**ncestries. As the name suggests, it uses a hidden Markov model to capture the LD along a chromosome, as in the SABER method. However, while SABER uses a second-order model to do this, HAPAA is more sophisticated in that it represents possible emissions in a manner that models long-range correlations more accurately than fixed-order models which assume a greater extent of independence between nearby sites. Furthermore, HAPAA has been shown to outperform SABER in a simulation study of three-way admixture.

It requires panels of phased haplotypes from samples of the candidate ancestral populations, typed at a set of linked SNPs and a possibly admixed study individual whose genotypes are also known at these loci. The structure of the model is hierarchical, with the states belonging to population, individual and haplotype levels. That is, the hidden states correspond to one of the two haplotypes of a particular individual in one of the ancestral panels. In a sense, the model tries to construct the admixed sample

from some piecing together of the observed panel haplotypes. To account for mutations, genotyping error and to permit haplotypes that may exist in the populations but are not present in the panels, the authors allow the allele of the study individual to differ from that at the corresponding state with some small probability that will be learnt from training data.

Both *inter* and *intra*-population transitions take place as each move along the chain may be (1) from a panel individual of a particular population to one in a different panel, (2) to another individual in the same panel, (3) to the other haplotype within the same individual, or even (4) back to the same state. Switches between individuals are modeled as a Poisson process of rate τ per unit genetic distance d. The genetic distance d_i between consecutive loci i and $i+1$ is known and population-specific recombination rate parameters τ_p are also learned from training data. At the lowest level, transitions between the pair of haplotypes within the same diploid individual k of population p at locus i are described as occurring with probability

$$w_{pki} \cdot e^{-\tau_p d_i}$$

Alternatively, a transition may result in returning back to the same state with probability

$$(1 - w_{pki}) \cdot e^{-\tau_p d_i}$$

The probability of transitioning to a higher level, that is to another individual possibly of a different population, is then

$$1 - e^{-\tau_p d_i}$$

Estimates of the vector of phase switch probabilities $\boldsymbol{w}$ between successive loci are estimated from the phasing method when the ancestral panels of haplotypes are generated from their genotypes at the beginning of the analysis. Transitions at the indi-

vidual level, between population p and $p^{'}$ say, are governed by an admixture matrix $A(p, p^{'})$ and there is a uniform probability $\frac{1}{2n_{p^{'}}}$ of moving to each of the haplotypes in population $p^{'}$ for $n_{p^{'}}$ individuals that panel. As with the emission probabilities and recombination rate parameters, the admixture transition matrix is estimated via supervised learning by applying this model to samples of uniform and known ancestry and requiring that these haplotypes are modeled within their true population.

Once all parameters have been estimated, the posterior probability of each state may be computed from the forward and backward algorithms and, for each locus, the population whose total posterior probability is greatest is the called ancestry. In an attempt to smooth out ancestry chunks that are very small, the authors include a post-processing step based upon the genetic length of the smallest acceptable chunk size ϵ. In this step, they fix the ancestry at chunks that are longer than ϵ to be that of largest posterior probability. At chunks deemed too small they re-calculate the marginal posterior probabilities for states of ancestry the same as the preceding and following chunks, insisting the ancestral call is made between these two populations.

1.7.8 SWITCH

SWITCH (Sankararaman et al., 2008a) is a probabilistic model whose HMM/MHMM framework features recombination event indicators, estimating the location of recombination events, where those that result in a change of ancestry are a subset of all recombination events that have occurred. The MHMM accounts for background LD in a manner similar to that of SABER. In calculating the emission probabilities, the SWITCH methods condition on whether or not any recombination event has occurred at all between the neighboring SNPs with an indicator variable $W_{i,j}$ to denote whether or not recombination occurred between SNPs $j-1$ and j on haplotype number i. It does this rather than conditioning on the ancestral state as is done in SABER which ig-

nores recombination events that do not result in a change in ancestry. These events will be common when the admixture consists of a small proportion q from one of the populations (i.e. a large proportion of the other). With an expected fraction of $q^2 + (1-q)^2$ of the recombination events *not* leading to a change in ancestry, the authors argue that these events should not be ignored.

Again, the occurrence of recombination is modeled as a Poisson process so

$$Pr(W_{i,j} = 1|g, r_j) = \theta_j$$

where

$$\theta_j = 1 - e^{-(g-1)d_j r_j}$$

and g denotes the number of generations of admixing, d_j is the distance between locus j and $j-1$, and r_j is the recombination rate at locus j. Unlike in the Structure linkage model and SABER, here the transition probabilities between ancestries at consecutive sites are also conditional on the variable W. If recombination does occur between between sites $j-1$ and j on haplotype i then the ancestry $Z_{i,j}$ will be 0, say, with probability equal to the global admixture fraction q or else it will be 1. On the other hand, if no recombination takes place then there can be no change in ancestry and $Z_{i,j}$ must equal $Z_{i,j-1}$. The transition probabilities used in the Structure linkage model and SABER may be obtained by marginalizing the SWITCH model transition probabilities over $W_{i,j}$.

The method can tackle two problems that require different inference, both assuming that the number of generations since the admixture is known. In the first case, the admixture fraction q is known but the ancestral frequencies are not, and the aim is to infer the local ancestries. In this scenario, they seek to find the maximum a posteriori (MAP) estimates of Z and W (where W is the indicator of whether there has been any recombination events or not) such that they maximize $log[Pr(W, Z|X, q, r)]$ where

X is the haplotype data. An application of the EM algorithm is used and the Viterbi algorithm finds the updates of Z and W.

The second problem concerns when the allele frequencies are known for one ancestral population only, and the goal is to find allele frequencies in the other population as well as q. If P_2 represents the unknown allele frequencies (those of population 2) then the method estimates

$$\arg\max_{P_2,q} logPr(P_2, q|X, P_1, r) = \arg\max_{P_2,q} logPr(X|P_1, P_2, q, r)$$

since uniform priors are placed on P_2 and q.

The Markov HMM (MHMM) used models background LD as a first-order Markov chain where emission probabilities are conditional upon not only the current ancestry state but also the haplotype at the preceding locus if no recombination has occurred otherwise they are, as in the HMM, conditional upon the ancestry state at the SNP only. If no recombination has occurred, the emission probability is the relevant joint allele frequency at the two SNPs. The model used is the same as that of SABER but the implementation differs as here ancestry estimates are computed via the Viterbi algorithm.

1.7.9 WINPOP

WINPOP (Pasaniuc et al., 2009) is a further development to LAMP that improves the modeling of recombination events using an adaptive window size. The ancestral allele frequencies $\{\mathbf{p}_1, \ldots, \mathbf{p}_K\}$ for each of the K populations are assumed to be known, as are the admixture fractions $\{q_1, \ldots, q_K\}$, and the method is applied to uncorrelated SNPs as it does not take in to account LD. In LAMP, the model assumes that no recombination events occur within a window, while WINPOP assumes at most one recent recombination event within each window.

At every position of the window along a chromosome, the optimal window length L is dependent upon the genetic distance between the two populations in that window. The estimate of L is initialized such that the probability of more than one recombination event occurring within the window is bounded by some small constant ϵ, such as 0.1. It is assumed that recombination events along a single chromosome are generated by a Poisson process of rate $(g-1)\rho$ where ρ is the recombination rate (assumed constant across the window) and g is the number of generations since the mixing began. It follow that the number of changes in ancestry occurs as a thinned version of this process with parameter $\lambda = 2(g-1)\rho \cdot (1 - \sum_{k=1}^{K} q_k^2)$ since $\sum_{k=1}^{K} q_k^2$ proportion of recombination events is expected to not result in a switch in ancestral state. Thus the probability the more than one switch occurs in the window is $1 - e^{-\lambda L} - \lambda L e^{-\lambda L} \approx (\lambda L)^2$ and the optimization of L is initialized at

$$L = \frac{\sqrt{\epsilon}}{2(g-1)\rho \cdot (1 - \sum_{k=1}^{K} q_k^2)}$$

An iterative search for the optimal value of L begins by either extending or shortening the window by 20 SNPs at a time, and checking if the new window length leads to a gain in accuracy when the model is tested on a simulated data set of admixed samples. The optimal value of L is therefore learnt from the simulations.

Once the optimal window lengths for each part of the chromosome have been found, then for each admixed individual the method seeks the site of recombination R and the two diploid population classifications, θ_1 and θ_2, of the upstream and downstream ancestry relative to the position of R. This is done by finding the maximum a posteriori estimates of (θ_1, θ_2, R) that maximize

$$P(\theta_1 = Z_{s_1} Z_{t_1}, \theta_2 = Z_{s_2} Z_{t_2}, R = r | \mathbf{p}_1, \ldots, \mathbf{p}_K, \mathbf{G}) \tag{1.7}$$

where $Z_{s_1}, Z_{t_1}, Z_{s_2}, Z_{t_2} \in \{1, \ldots, K\}$ are the ancestries of the SNPs upstream of the site of r and those downstream of r respectively. The probability function in expression (1.7) is proportional to

$$\begin{aligned}\propto & P(\mathbf{G}_1|\mathbf{p}_1, \ldots, \mathbf{p}_K, \theta_1 = Z_{s_1}Z_{t_1}) \cdot P(\mathbf{G}_2|\mathbf{p}_1, \ldots, \mathbf{p}_K, \theta_2 = Z_{s_2}Z_{t_2}) \\ & \cdot P(\theta_1 = Z_{s_1}Z_{t_1}) \cdot P(R = r) \cdot P(\theta_2 = Z_{s_2}Z_{t_2}|\theta_1 = Z_{s_1}Z_{t_1}, R = r)\end{aligned}$$

Assuming Hardy-Weinberg equilibrium within the admixed samples, the first two terms are estimated to be the product of the relevant allele frequencies for the sample's genotype within the window. The prior probability on the ancestry classification θ_1 is the product of the two admixture fractions $2^{1-\delta_{s_1t_1}} \cdot q_{s_1}q_{t_1}$ where $\delta_{s_1t_1} = 1$ if $s_1 = t_1$ and is 0 otherwise, which takes in to account that for heterogeneous ancestry there are two different ways it may occur. The probability that recombination happens between SNPs r and $r + 1$ is

$$P(R = r) = (1 - e^{-2(g-1)L\rho}) \cdot \frac{d_{r,r+1}}{L}$$

where $d_{r,r+1}$ is the distance in base pairs between the two SNPs and L is the window length, also in base pairs. The final term in expression (1.7) is the transition probability from the upstream ancestry to the downstream ancestry given the location of R. Assuming that the recombination happens on the chromosome with ancestries t_1 and t_2 then this probability is equal to $\frac{1}{2}q_{t_2}$. The posterior probabilities of these estimates are then compared to those obtained when it is assumed that no recombination events occur at all within the window, and the parameters that maximize the greater of the two probabilities are selected.

1.7.10 HAPMIX

The HAPMIX (Price et al., 2009) population genetic model makes use of a dense set of genome-wide data, extending the model of Li and Stephens (2003) to infer local ancestry in individuals who are admixed from two populations. The method requires a panel of phased individuals from two ancestral populations that should be closely related to the actual ancestral populations of the individuals under study. It views the admixed individuals' haplotypes as being sampled from the haplotypes of the panels, and calculates the likelihood that the admixed haplotype is a better match to those in one ancestral population over the other. A HMM is also employed in this method to incorporate information at neighboring loci.

A novel feature of HAPMIX is that in viewing admixed haplotypes to be a sample of ancestral haplotypes it includes a miscopying parameter (which may take on a different value for each population) to incorporate the probability that, in some regions, although the true ancestral population is population a it may be that the region is a better fit to the haplotypes of population b. In a sense, they are allowing copying from the wrong population to occur. The authors argue that since only a finite sample of haplotypes from the ancestral populations are available, the admixed individual may have a deep coalescence time with its ancestor in the true ancestral population and so may actually coalesce first with an ancestor in the other population. They report that taking in to account the possibility of incomplete lineage sorting in this manner greatly improves model performance and reduces spurious changes in ancestry, particularly when the two ancestral populations are closely related.

In the copying from ancestral haplotypes that creates an admixed haplotype, they consider that, at recombination points, switches between the haplotype being copied will either occur more recently than the time of admixture as part of the ancestry switching process, or they will have occurred prior to the admixture as part of the

within population switching process. Both of these switching processes are modeled as Poisson processes but of different rates. The emission probabilities allow for mutations to occur, the probability of which is dependent on whether miscopying occurs at the locus or not. The EM algorithm is used to update the parameters of the model to maximize the expected log-likelihood of the data.

While the model may be extended to apply to multiple populations, the current implementation of HAPMIX can only handle two-way admixture analysis. Due to this limitation, the authors have adopted a post-hoc approach to dealing with three-way admixture[1], however the HAPMIX software does not currently include this extension and so is limited to the analysis of two-way admixture. They first group together two of the three ancestral panels and use this as a single panel with the other panel being that of the third population on its own. The model is then run and estimates of local ancestry for population 1 and 2 combined, and of population 3 are obtained. The method is then run again, this time combining the another pair of populations together in a single panel, say populations 1 and 3, giving estimates of the proportion of ancestry from populations 1 and 3 collectively and also from population 2. In the third run, the last remaining pair of populations, 2 and 3, are combined together in one larger panel and the method is run to infer ancestry from these populations collectively versus that of population 1. At the end of these three runs, an estimate of the proportion of ancestry from each population at every locus is obtained through a least squares solution.

1.7.11 ADMIXTURE

ADMIXTURE (Alexander et al., 2009) offers a faster implementation of the likelihood model underlying STRUCTURE to estimate global ancestry Q and ancestral allele frequencies P. As with FRAPPE, it is a frequentist version of the Bayesian STRUCTURE method, where rather than sampling from the posterior distribution they instead max-

[1]Personal communication with Simon Myers.

imize the likelihood to obtain point estimates. Since the I individuals are considered independent and J SNPs are thinned to be unlinked, the log-likelihood function for a model of K ancestral populations is

$$\ell(Q,P) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(g_{ij}\cdot log\left(\sum_{k=1}^{K} q_{ik}p_{kj}\right) + (2-g_{ij})\cdot log\left(\sum_{k=1}^{K} q_{ik}(1-p_{kj})\right)\right) + C \quad (1.8)$$

where $g_{ij} \in \{0,1,2\}$ is the genotype of individual i at SNP j, where C is a constant.

They use a block relaxation algorithm (Leeuw, 1994) as their optimization technique which they find converges faster than the EM algorithm of FRAPPE, requiring tens rather than thousands of iterations, and provides more accurate parameter estimates. It involves finding the increment Δ in the parameter being updated that optimizes the second-order Taylor expansion of the $\ell(Q,P)$ approximation

$$\ell(Q,P) \approx \ell(Q^n,P^n) + d\ell(Q^n,P^n)\cdot\Delta + \frac{1}{2}\Delta^t d^2\ell(Q^n,P^n)\Delta$$

subject to the following constraints: $q_{ik} \geq 0$, $\sum_{k=1}^{K} q_{ik} = 1$ and $0 \leq p_{kj} \leq 1$. It alternates between the updating of Q while P is fixed, and updating P for fixed Q.

The partial derivatives of the log-likelihood in eq.(1.8) are straight forward

$$\frac{\partial\ell}{\partial q_{ik}} = \sum_{j=1}^{J}\left(g_{ij}\cdot\frac{p_{kj}}{\sum_{l=1}^{K} q_{il}p_{lj}} + (2-g_{ij})\cdot\frac{1-p_{kj}}{\sum_{l=1}^{K} q_{il}(1-p_{lj})}\right)$$

$$\frac{\partial\ell}{\partial p_{kj}} = \sum_{j=1}^{J}\left(g_{ij}\cdot\frac{q_{ik}}{\sum_{l=1}^{K} q_{il}p_{lj}} + (2-g_{ij})\cdot\frac{-q_{ik}}{\sum_{l=1}^{K} q_{il}(1-p_{lj})}\right)$$

and the relevant second-derivatives are

$$\frac{\partial^2 \ell}{\partial q_{ik} q_{im}} = \sum_{j=1}^{J} \left(g_{ij} \cdot \frac{p_{kj} p_{mj}}{\left(\sum_{l=1}^{K} q_{il} p_{lj} \right)^2} + (2 - g_{ij}) \cdot \frac{(1 - p_{kj})(1 - p_{mj})}{\left(\sum_{l=1}^{K} q_{il} (1 - p_{lj}) \right)^2} \right)$$

$$\frac{\partial^2 \ell}{\partial p_{kj} p_{mj}} = \sum_{j=1}^{J} \left(g_{ij} \cdot \frac{q_{ik} q_{im}}{\left(\sum_{l=1}^{K} q_{il} p_{lj} \right)^2} + (2 - g_{ij}) \cdot \frac{q_{ik} q_{im}}{\left(\sum_{l=1}^{K} q_{il} (1 - p_{lj}) \right)^2} \right)$$

where

$$\frac{\partial^2 \ell}{\partial q_{i_1 k} q_{i_2 k}} = 0 \text{ if } i_1 \neq i_2 \qquad \text{and} \qquad \frac{\partial^2 \ell}{\partial p_{kj_1} p_{kj_2}} = 0 \text{ if } j_1 \neq j_2$$

1.7.12 RFmix

Principal Component Analysis (PCA) is a multivariate analysis technique that constructs low-dimensional projections of the data that explain as much of the variance between samples, under the constraint that the PCs are orthogonal. PCA was first used to study human genetic variation by Luca Cavalli-Sforza and colleagues (Menozzi et al., 1978) and since then, with the appearance of larger data sets (Nelson et al., 2008) including more populations at vastly more markers, it has proved an effective tool in revealing population structure within Europe (Novembre et al., 2008) and India (Reich et al., 2009). RFmix (Bryc et al., 2010a) incorporates PCA into a HMM to infer ancestry. The hidden states of the model are the number of alleles inherited from, say, population 1 at each biallelic locus.

In this method, PCA is run on the normalized and scaled genotypes of both the admixed samples and the putative ancestral populations. The method then proceeds by considering non-overlapping windows of around 10-20 SNPs in turn. For each individual i at window w, the score s_{iw} at the SNPs within this window is found, that

is

$$s_{iw} = G_{iw} \times e_w$$

where G_{iw} is the matrix of the individual i's genotypes at window w and e_w is the corresponding vector of loadings. It is then assumed, for each window, that the scores of the ancestral individuals are normally distributed where the sample means and variances are taken as the estimated parameters of the normal distribution for each population. Given these distributions on the scores for each ancestry, the relative densities of the scores of the admixed samples can then be found and converted to probabilities. These probabilities then serve as the emission probabilities of the HMM.

The transition probabilities of the HMM are the same as those of the Structure linkage model, with the number of generations since admixture and the global ancestral proportions being specified rather than estimated however. The Viterbi algorithm is used to find the most likely path of local ancestry states for each admixed individual, giving an estimate of the number of alleles (0, 1 or 2) inherited from each population.

1.7.13 LAMP-LD

The LAMP-LD (Baran et al.) method is a further development of LAMP which uses an approximation of the Li and Stephens model (Li and Stephens, 2003) to capture the haplotype structure of the ancestral populations in a fixed-size state space that is smaller than the number of haplotypes in the ancestral panels, while ensuring the algorithm is fast and computationally efficient. Unlike HAPMIX and HAPAA which employ HMMs that are quadratic in the number of ancestral haplotypes within the panels, LAMP-LD reduces the size of the state space making it easier to handle large ancestral data sets and an order of magnitude faster to run than HAPMIX for example.

The LAMP-LD model considers that SNPs are grouped in consecutive non-overlapping windows of L SNPs and it begins by assuming that the ancestry within any window

is exclusively from one population. That is, it initially makes the assumption that no switches in ancestry occur within a window, although this is later relaxed to refine the location of switch boundaries. There are two levels of the HMM in the LAMP-LD approach. One operates, for each population of ancestry, between the L SNPs of each window at which there are some number of "prototypical " states S smaller than the number of haplotypes in the ancestral panel. By choosing S to be less than the number of ancestral haplotypes, the complexity of the model is reduced and the variation seen among the ancestral haplotypes is approximated by these representative states. Within a particular window, any haplotype may arise from any underlying hidden path composed of some combination of the S states across the L SNPs.

This within-window HMM describes the probability of the admixed sample's haplotype $H = H_1 H_2 \cdots H_L$ conditional upon the ancestry classification of that window. That is, within a single window and for each possible ancestry at that window we have

$$P(H|\tau, \epsilon, Q) = \sum_{\pi} \tau_0(s_0, \pi_1)\epsilon_1(\pi_1, H_1) \prod_{i=2}^{L} \tau_i(\pi_{i-1}, \pi_i)\epsilon(\pi_i, H_i) \tag{1.9}$$

where the sum is taken over all possible paths π across the window and Q denotes the set of states across the L SNPs. The transition probabilities τ and emission probabilities ϵ of this model are estimated directly for each candidate ancestral population from the ancestral haplotypes themselves using the Baum-Welch algorithm.

The other level of the HMM describes how the ancestry varies from window to window. There is an emission probability associated with the sample's genotype G_w at each window given its diploid ancestral state (Z_1^w, Z_2^w). This probability is

$$\sum_{(H_1^w, H_2^w)} P(H_1^w|Z_1^w) \cdot P(H_2^w|Z_2^w)$$

where the sum is taken over all pairs of haplotypes (H_1^w, H_2^w) that are compatible with

the observed genotype G_w. The transition probabilities of changes in ancestral states from window to window are set to be constants dependent on how many of the two ancestries in the diploid state have changed in the transition. Specifically, if (Z_1^w, Z_2^w) and (Z_1^{w+1}, Z_2^{w+1}) differ by one ancestry then the transition probability is $\theta = D \times 10^{-8}$ where D is the number of base pairs between windows w and $w+1$. It is θ^2 if both ancestries differ and $1 - 2\theta - 3\theta^2$ if there are no changes in the ancestry pairs.

As in the original formulation of LAMP, the use of non-overlapping windows means that ancestral switches can occur at the end points of windows only. In LAMP-LD however this restriction is then relaxed in a post-processing stage which searches for the optimal location of the breakpoint across an interval around the end points of the windows where a change in ancestry was inferred. Using the within-window HMM of the two ancestral populations involved in the switch, the probability of the observed haplotype is computed for each possible location of the breakpoint within the search region and the site at which this likelihood is maximized is deemed the location of the switch.

1.8 Summary

As we have seen, there is a wide range of model-based (STRUCTURE, STRUCTURE 2, ANCESTRYMAP, FRAPPE, SABER, HAPAA, SWITCH, HAPMIX, ADMIXTURE, LAMP-LD) and non-model-based (LAMP, WINPOP) methods to estimate ancestry, with the PCA HMM of Bryc et al. making use of both of these approaches.

Central to most of the model-based techniques (STRUCTURE 2, ANCESTRYMAP, SABER, HAPAA, SWITCH, HAPMIX, RFmix LAMP-LD) is a HMM to capture the correlations in ancestry that are expected between loci that lie close together on a chromosome. To be able to estimate ancestry at a dense set of markers, these methods should take in to account background LD; something that is done to varying extents in

SABER, HAPAA, SWITCH, HAPMIX and LAMP-LD. The most sophisticated model on LD is that of HAPMIX, but this comes at the cost of generality as HAPMIX is not applicable when the admixture involves more than two populations. LAMP-LD uses an approximation of the Li and Stephens model of haplotypic variation to reduce the number of possible states in the HMM to speed-up computation without much loss in accuracy, but as with HAPMIX it also requires that the ancestral panels are phased.

In the next chapter we will present a novel model-based method of ancestry estimation - MULTIMIX - that takes in to account background LD, does not require phasing of either the ancestral panels or the admixed samples, is applicable to multi-way admixture while remaining computationally fast and practical for genome-wide analysis.

Chapter 2

Novel Methods of Ancestry Estimation

We have seen in Chapter 1 that there are a series of statistical techniques in the literature that were developed to infer genetic ancestry, and that their applications extend to admixture mapping, case-control association studies, and historical and anthropological population genetic studies. Despite the array of methods proposed, there is not as yet one that is applicable to individuals of multi-way admixture, that can handle dense genome-wide data at linked loci and remain computationally fast. In this chapter we propose a new model to be used in estimating the ancestry of an individual that satisfies all of these criteria. We then describe three different approaches of fitting the model to the data and suggest a further step to refine estimates of the location of switches in ancestry.

2.1 Notation

The purpose of our method is to infer the unobservable local ancestry $\mathbf{Z}$ at a set of L biallelic SNPs along an individual's chromosome, which may be represented by a vector of ancestry states

$$\mathbf{Z} = (Z_1, Z_2, ..., Z_L)$$

where each component Z_i denotes the ancestral origin of the allele at SNP i for i in $1, ..., L$. Where the SNP set is dense and genome-wide, learning of these local ancestry states will provide a detailed picture of the individual's genetic ancestral origins. We stress that our method is applied to a single individual at a time and does not require multiple samples from the admixed population being studied. The global ancestral proportions $\mathbf{q} = (q_1, \ldots, q_K)$ may be calculated from these local ancestries as

$$q_k = \sum_{i=1}^{L} \frac{\mathbb{1}[Z_i = k]}{L}$$

for each population k, which are simply the proportions of the alleles at the SNPs being studied that have been inherited from each ancestral population. The study individual has been typed at these L SNPs and this genotype information may or may not have been converted to phased haplotypes. A number of statistical phasing algorithms are available to do this, including SHAPE-IT (Delaneau et al., 2011), MACH (Li et al., 2010), IMPUTE (Howie et al., 2009), and BEAGLE (Browning and Browning, 2007). As we will show, our method is applicable in both instances. Where the study individual is phased we have an observed haplotype $\mathbf{X}$ at the same set of L SNPs denoted by

$$\mathbf{X} = (X_1, X_2, \ldots, X_L)$$

where $X_l \in \{0, 1\}$ for $l \in \{1, \ldots, L\}$, and when it is unphased then $\mathbf{X}$ represents the genotype instead and $X_l \in \{0, 1, 2\}$.

Consider that we have a panel of haplotypes or genotypes from each of the K candidate ancestral populations, typed at the same set of L SNPs as the sample. We stress that, as with the study individual, our method is applicable when the panel is either phased or unphased. For population $k \in \{1, ..., K\}$ we have N_k source haplotypes or genotypes in the panel $\mathbf{H}^{(k)}$ where say the SNPs are arranged in columns and the

samples are arranged in rows. That is, the panel for population k may be denoted

$$\mathbf{H}^{(k)} = (\mathbf{h}_1^{(k)}, \mathbf{h}_2^{(k)}, \ldots, \mathbf{h}_{N_k}^{(k)})^T$$

where the source haplotypes or genotypes are

$$\mathbf{h}_j^{(k)} = (h_{j1}^{(k)}, h_{j2}^{(k)}, \ldots, h_{jL}^{(k)})$$

and either $h_{jl}^{(k)} \in \{0, 1\}$ for a phased panel or $h_{jl}^{(k)} \in \{0, 1, 2\}$ for a genotype panel, for $j \in \{1, \ldots, N_k\}$ and $l \in \{1, \ldots, L\}$. We will use the notation $\mathbf{H}$ with no superscript to refer to the complete set of ancestral panels $\{\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(K)}\}$. This is the context in which our model is applicable and throughout this chapter we will assume that the scenario described above applies.

2.2 Fast estimation of global ancestry from unlinked SNPs

The first and most simple model we consider uses an EM algorithm to estimate the global admixture proportions of the individual. Here, the observed data is $\mathbf{X} = \{X_{il}\}$ for $X_{il} \in \{0, 1\}$ where i is the SNP index and $l \in \{1, 2\}$ is the index of the allele in the pair at that SNP. The labeling of the alleles in the pair as allele 1 and allele 2 is irrelevant since it does not take in to account any phase information. Consequently, this model applies regardless of whether the study individual is phased or not.

This model is applied to unlinked SNPs, which in practice may be obtained from a denser set by pruning to remove those that exhibit correlations above some threshold. For some such subset (S) of L unlinked SNPs we calculate the observed allele frequencies for each of the K populations from the corresponding ancestral panel. These are simply the sample means of the panel haplotypes at each site which we use to estimate the true allele frequencies. In population k at SNP i these allele frequencies shall be de-

noted $p_k^{(i)}$, with the vector of allele frequencies for all SNPs being represented as $\mathbf{p}_k$ in population k . The model relates the ancestral panel information and the unobserved ancestry $\mathbf{Z}$ to the individual's haplotype $\mathbf{X}$:

$$P(\mathbf{Z}, \mathbf{q}|\mathbf{X}; \mathbf{H}) \quad \propto \quad P(\mathbf{X}|\mathbf{Z}; \mathbf{H}, \mathbf{q}) \cdot P(\mathbf{Z}, \mathbf{q}) \tag{2.1}$$

We assume an objective prior on the distribution of $\mathbf{q}$ if we have no reason to believe *a priori* that the study individual has inherited a larger proportion of ancestry from one population than another. The Dirichlet distribution is used for $K > 2$, simplifying to a beta distribution in the two population case:

$$P(\mathbf{Z}, \mathbf{q}) \quad = \quad P(\mathbf{Z}|\mathbf{q}) \underbrace{P(\mathbf{q})}_{\mathbf{q} \sim Dir(\alpha)} \tag{2.2}$$

where α is a vector of length K with each entry equal to some constant c making it an objective prior.

The SNP subset is assumed to be sufficiently sparse such that correlations in ancestry along a chromosome may be ignored, and the probability of a particular ancestry at a locus is dependent on nothing other than the global ancestry proportions:

$$P(\mathbf{Z}|\mathbf{q}) \quad = \quad \prod_{i \in S} \prod_{l=1}^{2} q_1^{\delta_{z_{il}1}} \cdot q_2^{\delta_{z_{il}2}} \cdots q_K^{\delta_{z_{il}K}} \tag{2.3}$$

where $\delta_{z_{il}k} = 1$ if $z_{il} = k$, and 0 otherwise. Here the product is taken over all loci in S and the two alleles that have been typed at each locus. Assuming independence across loci, the likelihood of the study individual's haplotype is simply the product across all

sites of the probability of the observed alleles in the ancestral population of the allele. That is

$$P(\mathbf{X}|\mathbf{Z};\mathbf{H},\mathbf{q}) \;=\; \prod_{i\in S}\prod_{l=1}^{2} P(X_{il}|z_{il};\mathbf{p}_1,\ldots,\mathbf{p}_K)$$

where at each locus for each of the two alleles

$$P(X_{il}|z_{il};p_1^{(i)},\ldots,p_K^{(i)}) \;=\; (p_{z_{il}}^{(i)})^{X_{il}}\cdot(1-p_{z_{il}}^{(i)})^{1-X_{il}}$$

We use an EM algorithm, detailed below, to estimate the mode of the posterior distribution $P(\mathbf{Z},\mathbf{q}|\mathbf{X};\mathbf{H})$. Rather than maximizing the log-likelihood of the unknown parameter $\mathbf{q}$ given the observed data X, we instead work with the expectation of the log-likelihood of the complete data $(\mathbf{X},\mathbf{Z})$. The algorithm is an iterative two-step scheme that (1) calculates the expectation of the complete data given the current parameter estimate $\mathbf{q}^*$ and then (2) updates $\mathbf{q}$ to be the value that maximizes this expression. The following exposition of its application follows the description of Bilmes (1998).

The expectation we want to maximize is taken over the set of all possible hidden states Ω

$$\mathbb{E}_{\mathbf{Z}|\mathbf{X},\mathbf{q}^*}\left[logP(\mathbf{X},\mathbf{Z}|\mathbf{q})\right] = \sum_{\mathbf{Z}\in\Omega}\Big(logP(\mathbf{X},\mathbf{Z}|\mathbf{q})\cdot P(\mathbf{Z}|\mathbf{X},\mathbf{q}^*)\Big)$$

Taking the conditional probability of X given Z over all L loci this becomes

$$= \sum_{\mathbf{Z}\in\Omega}\left(\sum_{i=1}^{L}\sum_{l=1}^{2} log\left(P(\mathbf{x}_{il}|\mathbf{z}_{il})q_{z_{il}}\right)\cdot\prod_{j=1}^{L}\prod_{m=1}^{2}P(\mathbf{z}_{jm}|\mathbf{x}_{jm},\mathbf{q}^*)\right)$$

Expanding over all components of Z

$$= \sum_{n=1}^{2}\sum_{z_{1n}=1}^{K}\sum_{z_{2n}=1}^{K}\cdots\sum_{z_{Ln}=1}^{K}\left(\sum_{i=1}^{L}\sum_{l=1}^{2} log\left(P(\mathbf{x}_{il}|\mathbf{z}_{il})q_{z_{il}}\right)\cdot\prod_{j=1}^{L}\prod_{m=1}^{2}P(\mathbf{z}_{jm}|\mathbf{x}_{jm},\mathbf{q}^*)\right)$$

We use an indicator function δ to be able to write the log term independently of z_{il}

$$= \sum_{n=1}^{2} \sum_{z_{1n}=1}^{K} \sum_{z_{2n}=1}^{K} \cdots \sum_{z_{Ln}=1}^{K} \left(\sum_{i=1}^{L} \sum_{l=1}^{2} \sum_{\gamma=1}^{K} \delta_{\gamma z_{il}} log\left(P(\mathbf{x}_{il}|\gamma) q_\gamma\right) \cdot \prod_{j=1}^{L} \prod_{m=1}^{2} P(\mathbf{z}_{jm}|\mathbf{x}_{jm}, \mathbf{q}^*) \right)$$

allowing us to take this term out of the summation over the z_{il}

$$= \sum_{l=1}^{2} \sum_{i=1}^{L} \sum_{\gamma=1}^{K} log\left(P(\mathbf{x}_{il}|\gamma) q_\gamma\right) \underbrace{\sum_{n=1}^{2} \sum_{z_{1n}=1}^{K} \sum_{z_{2n}=1}^{K} \cdots \sum_{z_{Ln}=1}^{K} \delta_{\gamma z_{il}} \cdot \prod_{j=1}^{L} \prod_{m=1}^{2} P(\mathbf{z}_{jm}|\mathbf{x}_{jm}, \mathbf{q}^*)}_{P(\gamma|x_{il}, \mathbf{q}^*)}$$

$$= \sum_{l=1}^{2} \sum_{i=1}^{L} \sum_{\gamma=1}^{K} log\left(P(\mathbf{x}_{il}|\gamma) q_\gamma\right) \cdot P(\gamma|x_{il}, \mathbf{q}^*)$$

Expanding the log of the product gives

$$= \sum_{l=1}^{2} \sum_{i=1}^{L} \sum_{\gamma=1}^{K} log\left(P(\mathbf{x}_{il}|\gamma)\right) \cdot P(\gamma|x_{il}, \mathbf{q}^*) + \sum_{l=1}^{2} \sum_{i=1}^{L} \sum_{\gamma=1}^{K} log(q_\gamma) \cdot P(\gamma|x_{il}, \mathbf{q}^*)$$

The first of the two terms does not involve q_γ, but rather the current estimate $\mathbf{q}^*$, so we need only maximize the second term which must be done subject to the constraint that the global ancestry proportions sum to 1, that is

$$\sum_{\gamma=1}^{K} q_\gamma = 1$$

This constraint is introduced via a Lagrange multiplier λ and the equation to solve becomes

$$\frac{\partial}{\partial q_\gamma} \left[\sum_{l=1}^{2} \sum_{i=1}^{L} \sum_{\gamma=1}^{K} log(q_\gamma) \cdot P(\gamma|x_{il}, \mathbf{q}^*) + \lambda \left(\sum_{\gamma=1}^{K} q_\gamma - 1 \right) \right] = 0$$

$$\Leftrightarrow \sum_{l=1}^{2} \sum_{i=1}^{L} \frac{1}{q_\gamma} P(\gamma|x_{il}, \mathbf{q}^*) + \lambda = 0 \Leftrightarrow q_\gamma = \frac{-1}{\lambda} \sum_{l=1}^{2} \sum_{i=1}^{L} P(\gamma|x_{il}, \mathbf{q}^*)$$

Summing both sides over all values of γ gives

$$\sum_{\gamma=1}^{K} q_\gamma = \sum_{\gamma=1}^{K} \left[\frac{-1}{\lambda} \sum_{l=1}^{2} \sum_{i=1}^{L} P(\gamma | x_{il}, \mathbf{q}^*) \right] \Leftrightarrow 1 = \frac{-2L}{\lambda} \Leftrightarrow \lambda = -2L$$

Therefore the parameter update that maximizes the expectation of the complete log-likelihood is

$$q_\gamma = \frac{1}{2L} \sum_{l=1}^{2} \sum_{i=1}^{L} P(\gamma | x_{il}, \mathbf{q}^*)$$

Using Bayes' rule, at any locus we can compute the probability of a hidden state given the observed haplotype and the current parameter estimate as follows:

$$P(z_{il} = k | x_{il}; \mathbf{q}^*) = \frac{P(x_{il} | z_{il} = k) \cdot q_k^*}{P(x_{il})} = \frac{P(x_{il} | z_{il} = k) \cdot q_k^*}{\sum_{j=1}^{K} P(x_{il} | z_{il} = j) \cdot q_j^*}$$

Explicitly, the update is then

$$q_k = \frac{1}{2L} \sum_{l=1}^{2} \sum_{i=1}^{L} \left(\frac{(p_k^{(i)})^{x_{il}} \cdot (1 - p_k^{(i)})^{1 - x_{il}} \cdot q_k^*}{\sum_{j=1}^{K} P(x_{il} | z_{il} = j) \cdot q_j^*} \right)$$

This new estimate of q_k is then used in the expectation step and the algorithm is repeated until convergence.

This simple model is that underlying the methods STRUCTURE and FRAPPE except that here the allele frequencies are fixed and not parameters to be estimated. It provides a straight-forward approach to directly estimating the total proportion of ancestry inherited from each ancestral population.

2.3 Multivariate normal approximation in a HMM framework

We have just described a fast and accurate method to determine the global ancestral proportions of an individual that does not make use of phase information. Next we present a more sophisticated method, named MULTIMIX, that estimates local ancestry at a denser set of linked SNPs and utilizes the phase of the study individual if available.

As explained in section 1.2, the chromosomes of an admixed individual may be thought of as a series of segments consisting of alleles of common ancestry since those at loci that are near to each other tend to be inherited together during meiosis, with the sites of recombination events delineating the boundaries between segments of different ancestry. For realistic levels of recent admixture long stretches of loci will share the same ancestry as relatively few recombination events will have taken place since the admixture event. This suggests that we might be able to carry out inference at a slightly coarser scale than inferring ancestry one SNP at a time without any meaningful loss of accuracy. To do this we split each chromosome up into $W = ceiling(L/n)$ contiguous windows of at most n SNPs and denote S_w as the ancestry with the w^{th} window such that

$$S_w = s \rightarrow Z_i = s \text{ for } i \in [(w-1)n + 1, \ldots, min(wn, L)]$$

for SNP i.

Our aim is then to conduct inference on $\mathbf{S} = (S_1, S_2, ..., S_W)$, the ancestry classification of each window, as an approximation to Z, the ancestry at each locus. The model we use has two main components (a) a probabilistic model of how ancestry changes between windows along a chromosome, and (b) a statistical model for the distribution of the observed haplotypes within each window. These components are described in the following two subsections for the case of the study individual being phased, and

then in section 2.3.3 their extension to an unphased individual is explained.

2.3.1 Modeling ancestry within a window

Within a given window we need to formulate a model that is able to accurately discriminate the population of origin of an observed haplotype. Due to differences in allele frequencies between populations an observed haplotype will tend to be more likely from one source population than from others and our model must capture this property. Allele frequencies and levels of LD between subsets of SNPs vary between populations and so when using dense SNP data from a genotyping chip we need to take account of these factors. Since the data at each SNP is binary, we need a discrete multivariate distribution to do this and this might be best achieved using a coalescent model but it is challenging to conduct computationally efficient inference using these models. Our motivation is to take a more computationally tractable approach to capturing the variation within a window across different ancestral populations.

We do this using a multivariate normal distribution. The multivariate normal has been shown to be useful in other statistical methods for human genetics. Wen and Stephens (2010) have used it to impute allele frequencies at untyped SNPs in a study sample given those observed at typed SNPs combined with information on the correlations between typed and untyped SNPs in a larger reference panel. In this case, a multivariate normal distribution is employed to model allele frequencies across linked SNPs, rather than modeling the haplotypes themselves as in our model described here.

We use a simple model for the distribution of the observed haplotype $\mathbf{X}_w$ given a *fitted* ancestry Y_w

$$\mathbf{X}_w | Y_w \sim N(\boldsymbol{\alpha}_{Y_w}^{(w)}, \Sigma_{Y_w}^{(w)} + \lambda I_n) \tag{2.4}$$

where $\alpha_k^{(j)}$ and $\Sigma_k^{(j)}$ are the mean allele frequencies and covariance matrix of SNPs within the jth window from the kth population.

The parameters of the multivariate normal (MVN) distributions within each window are estimated directly from the source population haplotypes. Since these parameters are fixed values in the model, this need only be done once at the start of the inference and so is computationally efficient. The allele frequencies at each SNP are estimated to be the sample means observed in the corresponding panel adjusted such that the allele frequencies are never allowed to be exactly equal to 0 or 1. This allows the model to account for the possibility of haplotypes that have not been observed in the ancestral panels. In practice this is done by setting the estimate in population k at SNP l, say, to be

$$\frac{\sum_{i=1}^{N_k} \delta_{1X_{li}} + 0.1}{N_k + 0.2}$$

where X_{li} is the haplotype of the ith sample in the panel of population k at SNP l. This equates to assuming that a single heterozygote is seen for every $10N_k$ haplotypes observed, among which the allele occurs with the same frequency as found in the set of N_k panel haplotypes.

The estimate of the covariance matrix that we use is composed of the unbiased sample covariances with a modification such that a small positive value λ is added to the variances to ensure the matrix is positive semi-definite and therefore invertible and a valid covariance matrix. As in the adjustment of the sample means, this modification of the sample variances allows for the occurrence of haplotypes that were not sampled in the ancestral panels.

These parameters can also be estimated if the source population data is unphased. Allele frequencies are easily estimated from unphased data. To estimate the haplotypic covariance matrix we use half the genotypic covariance matrix, assuming that the genotype at any SNP is the sum of two independent identically distributed haplotypes and thus the covariance between haplotypes at different SNPs is zero if they are not on the same chromosome copy.

An alternative way to estimate the covariance matrices in the case of an unphased panel of genotypes would be to instead infer the frequency of the underlying two-locus haplotypes at pairs of SNPs. This may be done using an E.M. algorithm to estimate the frequency of the unobservable phased haplotypes, ambiguous due to samples that are heterozygous at both SNPs (Weir, 1996). This approach was investigated, however we found that it did not lead to a valid covariance matrix since the pairwise calculation of covariances did not necessarily result in a congruous structure of the matrix. The method described above, of halving the genotypic covariances, provides a straightforward and accurate estimate of the haplotypic covariances.

This model attempts to characterise the first two moments of the multivariate discrete distribution of haplotypes from a population within a given window. We do not claim that the use of a continuous probability model to model the discrete distribution of haplotypes provides a good *absolute* approximation to probabilities of observing a given haplotype from a particular population. We will show however that it can provide a *useful* approximation. To infer ancestry within a window, and subsequently across a whole chromosome, the method must be able to distinguish between several competing models of ancestry. For example, in the case of two source populations we will make correct inferences of ancestry if we can accurately model the ratio $P(\mathbf{X}_w|Y_w = a)/P(\mathbf{X}_w|Y_w = b)$, for two candidate fitted ancestries a and b. We stress that it is the *relative* probability of haplotypes given two competing models of ancestry that is important. In addition, the MVN model accounts for dependencies between SNPs within the same window but not between SNPs from different windows. In general, ignoring such dependencies can lead to errors and over confidence in statistical estimation but our aim here is to account for enough of the dependence that any bias is minimized. As will be seen in Chapter 3 the use of the multivariate normal model works very well in this regard.

2.3.2 Modeling changes in ancestry along a chromosome

As the STRUCTURE linkage model, ANCESTRYMAP, SABER, HAPAA, SWITCH, HAPMIX and PCA-HMM methods have done previously, we model switches in ancestry along a chromosome as a Markov process. We assume that recombination events along a chromosome occur as a Poisson process of rate r per unit genetic length, where only some of these events result in a switch in the ancestral state. Under the assumption of a single admixture event involving all of the ancestral populations, and assuming that after this event individuals of subsequent generations continue to mix freely then $(r+1)$ may be interpreted as the number of generations since the admixture occured. Assuming a first-order model, we model the joint distribution of S as

$$P(\mathbf{S}) = \prod_{w=1}^{W-1} P(S_{w+1} = i | S_w = j; \mathbf{q}, r) \tag{2.5}$$

where at the first window the states are initialized with probabilities equal to the global ancestry proportions

$$P(S_1 = k) = q_k \tag{2.6}$$

The transition probabilities between states in consecutive windows, S_w and S_{w+1}, are modeled as

$$\tau_{w_{ji}} := P(S_{w+1} = i | S_w = j; \mathbf{q}, r) = \delta_{ij} \cdot e^{-d_w r} + (1 - e^{-d_w r}) q_i \tag{2.7}$$

We use $\tau_{w_{ji}}$ to denote the transition probability between the state of underlying ancestry of population j at window w to that of underlying ancestry of population i at window $w+1$. Here d_w is the genetic distance, in Morgans, between the midpoints of the two windows and $\mathbf{q} = \{q_1, \ldots, q_K\}$ is the unobserved ancestry proportions of the individual. That is, between two consecutive windows w and $w+1$, the probability that no recombination takes place is $e^{-d_w r}$ and the probability that some number of re-

combination events take place is $1 - e^{-d_w r}$ after which the population that the ancestry switches to is drawn from the probability vector of global ancestry proportions $\mathbf{q}$.

At this point, we make a distinction between the *true* ancestral population S_w and the population Y_w to which we *fit* our MVN model within the wth window. We found that formulating the hidden states in this way and allowing Y_w to differ from S_w helped to avoid inferring spurious switches in ancestry. States in which Y_w is not equal to S_w are referred to as *misfitting* states. This misfitting property is analogous to the mis-copying process as presented by Price at el. in the HAPMIX model. They describe how an admixed haplotype can be modeled as a mosaic of copies of several ancestral haplotypes from different populations, where at some loci there is misfitting when the ancestral haplotype being copied is of a different population to the true underlying ancestry of the admixed sample at that site.

We assume that fitted population states in consecutive windows are conditionally independent given the true population states and are homogeneous within windows, that is

$$P(\mathbf{Y}|\mathbf{S}) = \prod_{w=1}^{W} P(Y_w|S_w)$$

where

$$P(Y_w = j|S_w = i) = m_{ij} \qquad i, j \in \{1, \ldots, K\}. \tag{2.8}$$

and $\sum_{j=1}^{K} m_{ij} = 1$. The matrix of misfitting probabilities $\mathbf{M} = \{m_{ij}\}$ can either be fixed or estimated from the data. This definition of the *fitted* ancestral states allows us to define the conditional distribution of the $\mathbf{X}$ as

$$P(\mathbf{X}|\mathbf{Y}) = \prod_{w=1}^{W} P(\mathbf{X}_w|Y_w).$$

With the above assumptions and choice of statistical models to describe the relationship between the observed haplotypes and the unknown ancestry states, our pro-

posed model fits in to a hidden Markov model framework. A benefit to this is that there are standard computational algorithms that facilitate the calculation of key probabilities that feature in the model fitting process and greatly decrease the computational complexity of the fitting. In the following section we will describe three different techniques to fit the model - Markov chain Monte Carlo (MCMC) sampling, the Expectation-Maximization (EM) algorithm and its variant the Classification-EM algorithm.

2.3.3 Extension of the model to an unphased study individual

One of the strengths of our MULTIMIX model is that it does not require phase information of the study individual. It may be directly extended to apply to an unphased individual. In this case, the observed data $\mathbf{X}_w$ denotes the genotype within a window rather than the haplotype. The true ancestral state at each window will now consist of two *unordered* components (S_{w_1}, S_{w_2}) corresponding to the population(s) of the true ancestry. Will we use the notation $\mathbf{S}^*_w$ to denote the two-component true ancestry. Similarly the fitted ancestral state will be the pair (Y_{w_1}, Y_{w_2}) which will be denoted $\mathbf{Y}^*_w$. That is, if $K = 2$ then there are three possible underlying ancestries: $(1,1), (1,2), (2,2)$, each of which we will represent by a single index number such that $\mathbf{S}^*_w = i^*$ where $i^* \in \{1,2,3\}$. These are also the possible diploid fitted states, that is $\mathbf{Y}^*_w = j^*$ for $j^* \in \{1,2,3\}$. In general, for the model of K ancestral populations there will be $n_K = \sum_{k=1}^{K} k$ diploid ancestry states since the ordering of the populations in the pair is irrelevant.

In the case of an unphased sample, our model on the distribution of the observed data is extended from eq.2.4 to

$$\mathbf{X}_w|(Y_{w_1}, Y_{w_2}) \sim N(\boldsymbol{\alpha}^{(w)}_{Y_{w_1}} + \boldsymbol{\alpha}^{(w)}_{Y_{w_2}}, \Sigma^{(w)}_{Y_{w_1}} + \Sigma^{(w)}_{Y_{w_2}} + 2\lambda I_n) \tag{2.9}$$

Consider that at window w the diploid underlying ancestral state is $l^* = (l_1, l_2)$ and that at the following window the underlying state is $i^* = (i_1, i_2)$ and the fitted ancestry state is $j^* = (j_1, j_2)$. The transition probabilities between the two-component underlying ancestral states is then

$$P\Big(\mathbf{S}^*_{w+1} = i^* | \mathbf{S}^*_w = l^*; \mathbf{q}, r\Big)$$

$$= P\Big(S_{w+1_1} = i_1, S_{w+1_2} = i_2 | S_{w_1} = l_1, S_{w_2} = l_2; \mathbf{q}, r\Big)$$

$$= \Big(\delta_{i_1 l_1} \cdot e^{-d_w r} + (1 - e^{-d_w r}) q_{i_1}\Big)\Big(\delta_{i_2 l_2} \cdot e^{-d_w r} + (1 - e^{-d_w r}) q_{i_2}\Big)$$

$$+(1 - \delta_{i_1 i_2})\Big(\delta_{i_1 l_2} \cdot e^{-d_w r} + (1 - e^{-d_w r}) q_{i_1}\Big)\Big(\delta_{i_2 l_1} \cdot e^{-d_w r} + (1 - e^{-d_w r}) q_{i_2}\Big)$$

$$= \tau_{w_{l_1 i_1}} \cdot \tau_{w_{l_2 i_2}} + (1 - \delta_{i_1 i_2}) \cdot \tau_{w_{l_2 i_1}} \cdot \tau_{w_{l_1 i_2}}$$

$$:= \tau^g_{w_{l^* i^*}} \tag{2.10}$$

for $i_1, i_2, l_1, l_2 \in \{1, \ldots, K\}$ and $i^*, j^* \in \{1, \ldots, n_K\}$. The second term in the sum arises because if the true diploid ancestry at window $w + 1$ is heterogeneous, that is if $i_1 \neq i_2$, then it may be that either the model switches from S_{w_1} to S_{w+1_1} along one chromosome copy and from S_{w_2} to S_{w+1_2} on the other, or that alternatively it switches from S_{w_1} to S_{w+1_2} along one and from S_{w_2} to S_{w+1_1} along the other.

The misfitting probabilities also take in to account that the pair of ancestries at any window is unordered so eq.2.8 is extended to

$$P\Big(\mathbf{Y}^*_w = j^* | \mathbf{S}^*_w = i^*\Big)$$

$$= P\Big(Y_{w_1} = j_1, Y_{w_2} = j_2 | S_{w_1} = i_1, S_{w_2} = i_2\Big)$$

$$= m_{i_1 j_1} m_{i_2 j_2} + (1 - \delta_{j_1 j_2}) \cdot m_{i_1 j_2} m_{i_2 j_1}$$

$$:= m^g_{i^* j^*} \tag{2.11}$$

where the second term in the sum features if the fitted ancestry (Y_{w_1}, Y_{w_2}) is heterogeneous.

Equations 2.9, 2.10 and 2.11 describe the extension to an unphased study individual for the model of ancestry within a window and the model of how ancestry changes along a chromosome. With this addition, we now have a model that can be used to estimate local ancestry from either haplotype or genotype data of an admixed individual. In the following section we show how this MULTIMIX model may be fit to the data to carry out this ancestry inference.

2.4 Techniques of Model Fitting

We will now explain how three different statistical methods may be used to fit our model : MCMC sampling, an EM-algorithm and a Classification-EM (CEM) algorithm. We present each method in turn, describing its application to a phased study individual. We also give the extension of the MCMC method to the case where the individual is unphased. Each of these techniques offers a different means of performing inference on the ancestry of the study individual. In the MCMC scheme, the aim is to learn about the joint posterior distribution of hidden states and unknown parameters by drawing samples from it. In the EM-algorithm we indirectly maximize $P(\mathbf{X}|\mathbf{q}, r, \mathbf{M})$ by instead maximizing the expectation of the complete likelihood $P(\mathbf{X}, \mathbf{S}, \mathbf{Y}|\mathbf{q}, r, \mathbf{M})$ of the observed and hidden data over all possible ancestries. The CEM-algorithm, a variation on this, will find the sequence of ancestry states that are jointly most likely given the observed data.

2.4.1 MCMC sampling

The MCMC scheme described here is similar to that of the STRUCTURE linkage method (Falush et al., 2003). We begin by assuming that the study individual is phased then

in section 2.4.2 we specify the adaptations of the method to an unphased individual. Having observed the haplotype $\mathbf{X}$ of an admixed individual, the joint posterior distribution of the hidden ancestry states $(\mathbf{S}, \mathbf{Y})$ and the unknown parameters $\mathbf{q}$ and r is

$$P\left(\mathbf{S}, \mathbf{Y}, \mathbf{q}, r|\mathbf{X}, \mathbf{H}\right) \propto P\left(\mathbf{X}|\mathbf{Y}, \mathbf{H}\right) P(\mathbf{Y}|\mathbf{S}) P\left(\mathbf{S}|\mathbf{q}, r\right) P\left(\mathbf{q}\right) P(r)$$

We do not directly compute this distribution, but rather use MCMC sampling methods to draw independent random samples $(\mathbf{S}, \mathbf{Y}, \mathbf{q}, r)$ from it. A Markov chain is constructed such that its stationary distribution is $P\left(\mathbf{S}, \mathbf{Y}, \mathbf{q}, r|\mathbf{X}, \mathbf{H}\right)$ and we base our inference of the unknown parameters upon summary statistics of these samples. For some initial values of $\mathbf{q}$ and r, we start by calculating the forward probabilities

$$P\left(\mathbf{x}_1, \ldots, \mathbf{x}_w, \mathbf{S}_w = i, \mathbf{Y}_w = j|\mathbf{q}, r, \mathbf{H}\right) \tag{2.12}$$

denoted $\alpha_w^{(i,j)}$, for every possible ancestral state $(\mathbf{S}, \mathbf{Y}) = (i, j)$ at each window w in the HMM using standard techniques (Rabiner, 1989). For each chromosome, the forward probabilities at the first window, $\alpha_1^{(i,j)}$, are simply

$$P\left(\mathbf{x}_1, \mathbf{S}_1 = i, \mathbf{Y}_1 = j|\mathbf{q}, r, \mathbf{H^{(j)}}\right) = q_i \cdot m_{ij} \cdot p_{1_j}$$

where p_{1_j} is the emission probability of the sample haplotype in window 1 when fit to population j. The probabilities at all other windows are calculated sequentially as

$$\alpha_{w+1}^{(i,j)} = \sum_{u=1}^{K} \sum_{v=1}^{K} \alpha_w^{(u,v)} \cdot \tau_{w_{ui}} \cdot p_{w+1_j} \cdot m_{i,j} \tag{2.13}$$

where (i, j) is the state at window $w + 1$ and the summation is taken over all possible states (u, v) at window w. After one complete pass of the forward algorithm, we then begin the backward Gibbs sampling of (S, Y) starting with (S_W, Y_W), the ancestral state

at the final, or Wth, window since

$$\begin{aligned} P\left(S_W = i, Y_W = j | \mathbf{X}, \mathbf{q}, r, \mathbf{H}\right) &= \frac{P\left(S_W = i, Y_W = j, \mathbf{X} | \mathbf{q}, r, \mathbf{H}\right)}{P\left(\mathbf{X} | \mathbf{q}, r, \mathbf{H}\right)} \\ &\propto \alpha_W^{(i,j)} \end{aligned} \tag{2.14}$$

For preceding windows, the probability with which we sample the ancestral state at window w if we have sampled state (u, v) at window $w+1$ is

$$P\left(S_w = i, Y_w = j | S_{w+1} = u, Y_{w+1} = v, \ldots, S_W, Y_W, \mathbf{X}, \mathbf{q}, r, \mathbf{H}\right)$$

$$\propto P\left(S_w = i, Y_w = j, \mathbf{x}_1, \ldots, \mathbf{x}_w | \mathbf{q}, r, \mathbf{H}\right) P\left(S_{w+1} = u, Y_{w+1} = v, \ldots, S_W, Y_W, \mathbf{x}_{w+1}, \ldots, \mathbf{x}_W | \mathbf{q}, r, \mathbf{H}\right)$$

$$\propto \alpha_w^{(i,j)} \cdot \tau_{w_{iu}}$$

Once we have done this for every window along the chromosome, we have a sample $(\mathbf{s}, \mathbf{y})^{(t)}$ of the hidden ancestral states at iteration t and the next step is to perform Metropolis-Hastings updates of the parameters $\mathbf{q}$ and r. We choose a uniform prior on $\mathbf{q}$ and a Dirichlet proposal distribution with density function $f(\mathbf{q}^{(t+1)}; \mathbf{q}^{(t)})$ such that

$$\mathbf{q}^{(t+1)} \sim Dir(c \cdot \mathbf{q}^{(t)})$$

from which we draw a proposed update $\mathbf{q}^{(t+1)}$ dependent upon its current value $\mathbf{q}_t$. In practice, setting the factor c to be 100 was found to give good mixing of the chain with acceptance probabilities typically ranging from 77-85%. The Metropolis-Hastings ratio, $\phi_{\mathbf{q}}$, at iteration $t+1$ is

$$\phi_{\mathbf{q}} = \frac{P\Big((\mathbf{S}, \mathbf{Y}) = (\mathbf{s}, \mathbf{y})^{(t+1)} | \mathbf{q}^{(t+1)}, r\Big) \cdot P(\mathbf{q}^{(t+1)}) \cdot f(\mathbf{q}^{(t)}; \mathbf{q}^{(t+1)})}{P\Big((\mathbf{S}, \mathbf{Y}) = (\mathbf{s}, \mathbf{y})^{(t+1)} | \mathbf{q}^{(t)}, r\Big) \cdot P(\mathbf{q}^{(t)}) \cdot f(\mathbf{q}^{(t+1)}; \mathbf{q}^{(t)})}$$

$$= \frac{P\Big((\mathbf{S},\mathbf{Y}) = (\mathbf{s},\mathbf{y})^{(t+1)} | \mathbf{q}^{(t+1)}, r\Big) \cdot f(\mathbf{q}^{(t)}; \mathbf{q}^{(t+1)})}{P\Big((\mathbf{S},\mathbf{Y}) = (\mathbf{s},\mathbf{y})^{(t+1)} | \mathbf{q}^{(t)}, r\Big) \cdot f(\mathbf{q}^{(t+1)}; \mathbf{q}^{(t)})} \tag{2.15}$$

where the uniform priors on $\mathbf{q}$ cancel. The proposed update $\mathbf{q}^{(t+1)}$ is accepted with probability $min(1, \phi_{\mathbf{q}})$.

In the update of r, the proposal distribution is normal with mean equal to the current value of r, that is $r^{(t)}$,

$$r^{(t+1)} \sim N(r^{(t)}, \sigma^2)$$

where $\sigma^2 = 4$ and we bound r such that $0 \leq r \leq 100$. The density function is denoted $f(r^{(t+1)}; r^{(t)})$. We place a uniform prior on $r \sim U[0, 100]$ so the Metropolis-Hastings ratio is

$$\phi_r = \frac{P\Big((\mathbf{S},\mathbf{Y}) = (\mathbf{s},\mathbf{y})^{(t+1)} | \mathbf{q}, r^{(t+1)}\Big) \cdot P(r^{(t+1)}) \cdot f(r^{(t)}; r^{(t+1)})}{P\Big((\mathbf{S},\mathbf{Y}) = (\mathbf{s},\mathbf{y})^{(t+1)} | \mathbf{q}, r^{(t)}\Big) \cdot P(r^{(t)}) \cdot f(r^{(t+1)}; r^{(t)})}$$

$$= \frac{P\Big((\mathbf{S},\mathbf{Y}) = (\mathbf{s},\mathbf{y})^{(t+1)} | \mathbf{q}^{(t+1)}, r\Big)}{P\Big((\mathbf{S},\mathbf{Y}) = (\mathbf{s},\mathbf{y})^{(t+1)} | \mathbf{q}^{(t)}, r\Big)} \tag{2.16}$$

since the proposal and prior distributions cancel. As in the update of $\mathbf{q}$, the proposed value $r^{(t+1)}$ is accepted with probability $min(1, \phi_r)$.

At every iteration t, we obtain a sample $(\mathbf{s}, \mathbf{y}, \mathbf{q}, r)^{(t)}$ from the joint posterior distribution of the unknown ancestral states and parameters. At each window, the marginal posterior probability of each ancestry state (S_w, Y_w) is estimated by the proportion of times that state was sampled after the burn-in. The population of the underlying ancestry for which the sum of these probabilities over all fitted populations is largest will be taken to be the classification of the local ancestry. That is, we obtain marginal posterior probabilities of the ancestry for each window from the number of times each ancestral state has been sampled post-burn in and classify the inferred underlying ancestral

state $\hat{z_w}$ to be that which has the highest marginal posterior probability, that is

$$\hat{z_w} = \underset{k \in \{1,\ldots,K\}}{\arg\max} \sum_{k'=1}^{K} P(S_w = k, Y_w = k' | \mathbf{X}, \hat{\mathbf{q}}, \hat{r})$$

The sample means of the set of post-burn-in values sampled for $\mathbf{q}$ and r are the point estimates of these parameters.

2.4.2 Extension of the MCMC method to an unphased study individual

In section 2.3.3 we described how our model may be extended to apply to an unphased study individual where only genotype data is available. It requires that the technique of performing inference on the model must also be adapted in this case. Here we specify how the MCMC sampling scheme presented above is modified.

Recall that the true ancestral state at each window $\mathbf{S}^*_w = (S_{w_1}, S_{w_2})$ will now consist of two unordered components $i^* = (i_1, i_2)$ corresponding to the population(s) that have contributed to the ancestry at that window. Similarly the fitted state $\mathbf{Y}^*_w = (Y_{w_1}, Y_{w_2})$ will have two components $j^* = (j_1, j_2)$. It follows that the forward probabilities at the first window are

$$P\left(\mathbf{x}_1, S_{1_1} = i_1, S_{1_2} = i_2, Y_{1_1} = j_1, Y_{1_2} = j_2 | \mathbf{q}, r, \mathbf{H}^{(\mathbf{j_1})}, \mathbf{H}^{(\mathbf{j_2})}\right) = q^g_{i^*} \cdot m^g_{i^* j^*} \cdot p^g_{1_{j^*}}$$

where the superscript g is used to distinguish the terms from their haploid analogues. The term $q^g_{i^*}$ is the prior probability of the pair of ancestries, one being of population i_1 and the other of population i_2 at that window:

$$q^g_{i^*} = 2^{1-\delta_{i_1 i_2}} \cdot q_{i_1} q_{i_2}$$

The factor of 2 takes in to account that there are two ways a state of mixed ancestry can occur. The diploid misfitting probability $m^g_{i^*j^*}$ is expressed in terms of the relevant haploid misfitting probabilities as in eq.(2.11). The emission probability $p^g_{1_{j^*}}$ is computed from the log-density of the MVN distribution defined in eq.(2.9) for the individual's genotype within the window given the fitted ancestries are j_1 and j_2. As with the emission probabilities and the misfitting probabilities, the transition probabilities τ^g_w in the model for an unphased individual are derived from the haploid probabilities as in eq.(2.10). At all subsequent windows, the forward probabilities are

$$\alpha^{(i^*,j^*)}_{w+1} = \sum_{u^*=1}^{n_K} \sum_{v^*=1}^{n_K} \alpha^{(u^*,v^*)}_w \cdot \tau^g_{w_{u^*i^*}} \cdot p^g_{w+1_{j^*}} \cdot m^g_{i^*j^*}$$

We perform backward sampling of the diploid ancestry states where

$$P\left(\mathbf{S}^*_w = i^*, \mathbf{Y}^*_w = j^* | \mathbf{S}^*_{w+1} = u^*, \mathbf{Y}^*_{w+1} = v^*, \ldots, \mathbf{S}^*_W, \mathbf{Y}^*_W, \mathbf{X}, \mathbf{q}, r, \mathbf{H}\right)$$

$$\propto \alpha^{(i^*,j^*)}_w \cdot \tau^g_{w_{i^*u^*}}$$

The parameters $\mathbf{q}$ and r are updated via the Metropolis-Hastings algorithm as in the MCMC method for a phased study individual with the ratios of eq.(2.15) and eq.(2.16) respectively. After a sufficient number of samples post-burnin, we have a set of samples $\{\mathbf{S}^*_w\}$ and $\{\mathbf{Y}^*_w\}$ of the true and fitted diploid ancestries at each window. The marginal probabilities of each diploid state of true ancestries may be estimated by the proportion of samples that were drawn with that true ancestry state.

2.4.3 EM algorithm

The EM algorithm offers an alternative approach to fitting the model. Whereas MCMC sampling delivers a posterior distribution on the unknown parameters, the EM algorithm provides point estimates with the benefit of being extremely fast to converge.

The overall idea is to adjust the parameters $(\mathbf{q}, r, \mathbf{M})$ so as to maximize the probability of the observed data $\mathbf{X}$ given the model, that is $P(\mathbf{X}|\mathbf{q}, r, \mathbf{M})$. This is made tractable by introducing the hidden variables, or ancestry states, $\mathbf{Z}^*$ which we will use as shorthand notation for $(\mathbf{S}, \mathbf{Y})$. Rather than maximizing the incomplete-date likelihood $P(\mathbf{X}|\mathbf{q}, r, \mathbf{M})$, the aim is instead to find the parameters that maximize the expectation of the complete data likelihood $P(\mathbf{X}, \mathbf{Z}^*|\mathbf{q}, r, \mathbf{M})$ over $\mathbf{Z}^*$.

To see why this works, consider (in the haploid case) that at iteration $t+1$ we seek an update of the parameter estimates such that $logP(\mathbf{X}|\mathbf{q}, r, \mathbf{M}) > logP(\mathbf{X}|\mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)})$,

with the difference between these two log-likelihoods being

$$logP(\mathbf{X}|\mathbf{q},r,\mathbf{M}) - logP(\mathbf{X}|\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})$$

$$=log\left(\sum_{\mathbf{Z}^*} P(\mathbf{X}|\mathbf{Z}^*,\mathbf{q},r,\mathbf{M})P(\mathbf{Z}^*|\mathbf{q},r,\mathbf{M})\right) - logP(\mathbf{X}|\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})$$

$$=log\left(\sum_{\mathbf{Z}^*} \frac{P(\mathbf{X}|\mathbf{Z}^*,\mathbf{q},r,\mathbf{M})P(\mathbf{Z}^*|\mathbf{q},r,\mathbf{M})}{P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})} \cdot P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})\right)$$
$$- logP(\mathbf{X}|\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})$$

$$\geq \sum_{\mathbf{Z}^*} P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)}) \cdot log\left(\frac{P(\mathbf{X}|\mathbf{Z}^*,\mathbf{q},r,\mathbf{M})P(\mathbf{Z}^*|\mathbf{q},r,\mathbf{M})}{P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})}\right)$$
$$- logP(\mathbf{X}|\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)}) \text{ by Jensen's inequality}$$

$$= \sum_{\mathbf{Z}^*} P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)}) \cdot log\left(\frac{P(\mathbf{X}|\mathbf{Z}^*,\mathbf{q},r,\mathbf{M})P(\mathbf{Z}^*|\mathbf{q},r,\mathbf{M})}{P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})}\right)$$
$$- \sum_{\mathbf{Z}^*} P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)}) logP(\mathbf{X}|\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})$$

$$= \sum_{\mathbf{Z}^*} P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)}) \cdot log\left(\frac{P(\mathbf{X}|\mathbf{Z}^*,\mathbf{q},r,\mathbf{M})P(\mathbf{Z}^*|\mathbf{q},r,\mathbf{M})}{P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})P(\mathbf{X}|\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})}\right)$$

So we have found a lower bound for the log-likelihood $logP(\mathbf{X}|\mathbf{q},r,\mathbf{M})$:

$$logP(\mathbf{X}|\mathbf{q},r,\mathbf{M}) \geq logP(\mathbf{X}|\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)}) + \sum_{\mathbf{Z}^*} P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)}) \cdot log\left(\frac{P(\mathbf{X}|\mathbf{Z}^*,\mathbf{q},r,\mathbf{M})P(\mathbf{Z}^*|\mathbf{q},r,\mathbf{M})}{P(\mathbf{Z}^*|\mathbf{X},\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})P(\mathbf{X}|\mathbf{q}^{(t)},r^{(t)},\mathbf{M}^{(t)})}\right)$$

so increasing the term on the right hand side of the inequality ensures that the log-

likelihood of the observed data is also increased. We find the parameter updates $(\mathbf{q}^{(t+1)}, r^{(t+1)}, \mathbf{M}^{(t+1)})$ that maximize this term, that is

$$\begin{aligned}
&(\mathbf{q}^{(t+1)}, r^{(t+1)}, \mathbf{M}^{(t+1)}) \\
&= \underset{(\mathbf{q},r,\mathbf{M})}{\arg\max} \Bigg[logP(\mathbf{X}|\mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}) \\
&+ \sum_{\mathbf{Z}^*} P(\mathbf{Z}^*|\mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}) \cdot log \left(\frac{P(\mathbf{X}|\mathbf{Z}^*, \mathbf{q}, r, \mathbf{M}) P(\mathbf{Z}^*|\mathbf{q}, r, \mathbf{M})}{P(\mathbf{Z}^*|\mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}) P(\mathbf{X}|\mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)})} \right) \Bigg]
\end{aligned}$$

Retaining only the terms that vary with $(\mathbf{q}, r, \mathbf{M})$ this becomes:

$$\begin{aligned}
(\mathbf{q}^{(t+1)}, r^{(t+1)}, \mathbf{M}^{(t+1)}) &= \underset{(\mathbf{q},r,\mathbf{M})}{\arg\max} \left[\sum_{\mathbf{Z}^*} P(\mathbf{Z}^*|\mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}) \cdot log \left(P(\mathbf{X}|\mathbf{Z}^*, \mathbf{q}, r, \mathbf{M}) P(\mathbf{Z}^*|\mathbf{q}, r, \mathbf{M}) \right) \right] \\
&= \underset{(\mathbf{q},r,\mathbf{M})}{\arg\max} \left[\sum_{\mathbf{Z}^*} P(\mathbf{Z}^*|\mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}) \cdot logP(\mathbf{X}, \mathbf{Z}^*|\mathbf{q}, r, \mathbf{M}) \right] \\
&= \underset{(\mathbf{q},r,\mathbf{M})}{\arg\max}\, \mathbb{E}_{\mathbf{Z}^*} \left[logP(\mathbf{X}, \mathbf{Z}^*|\mathbf{q}, r, \mathbf{M}) \right]
\end{aligned}$$

Therefore, we seek the parameter values that maximize the expectation of the complete data log-likelihood.

In our model, the complete data consists of the underlying and fitted ancestry at each locus, the positions at which recombination events took place inclusive of those that did not result in a change in the underlying ancestry as well as those that did, and lastly the observed haplotype or genotype of the admixed sample across the region of the genome being considered. Let n_{ijuv} be the number of switches that occur from a window of ancestral state $(S_w = i, Y_w = j)$, that is where the underlying ancestry is population i and the fitted ancestry is population j, to one in which the ancestral state is $(S_{w+1} = u, Y_{w+1} = v)$ due to a recombination event taking place. We will use the dot notation for n_{ijuv} to denote the summing over a particular index, that is $n_{ij.v} = \sum_{u=1}^{K} n_{ijuv}$

for example. The likelihood of the complete data may now be written as:

$$P(\mathbf{X}, \mathbf{S}, \mathbf{Y}|\mathbf{q}, r, \mathbf{M}) = P(n_{....}) \cdot \prod_{k=1}^{K} q_k^{\mathbb{1}[S_1=k]} \cdot \prod_{k=1}^{K} q_k^{n_{..k.}} \cdot \prod_{w=1}^{W} \prod_{i=1}^{K} \prod_{j=1}^{K} m_{ij}^{\mathbb{1}[(S_w,Y_w)=(i,j)]} \cdot \prod_{w=1}^{W} P(\mathbf{X}_w|Y_w, \mathbf{H})$$

Since the emission probabilities $P(\mathbf{X}_w|Y_w, \mathbf{H})$ are independent of the parameters $(\mathbf{q}, r, \mathbf{M})$ that we wish to optimize, the log-likelihood may be written:

$$\ell(\mathbf{q}, r, \mathbf{M}|\mathbf{X}, \mathbf{Z}) = -rD + n_{....} log(r) + \sum_{k=1}^{K} \mathbb{1}[S_1 = k] \cdot log(q_k) + \sum_{k=1}^{K} n_{..k.} log(q_k)$$

$$+ \sum_{w=1}^{W} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbb{1}[(S_w, Y_w) = (i, j)] \cdot log(m_{ij}) + c \tag{2.17}$$

where c is a constant. We can now maximize the expectation of this log-likelihood where the expectation is taking with respect to the current set of parameter values $(\mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)})$ at iteration t.

First, taking the partial derivative of the expected log-likelihood with respect to r and setting this equal to zero to solve for the parameter update $r^{(t+1)}$:

$$\frac{\partial \mathbb{E}[\ell(\mathbf{q}, r, \mathbf{M}|\mathbf{X}, \mathbf{S}, \mathbf{Y})]}{\partial r} = -D + \frac{\mathbb{E}[n_{....}]}{r} = 0$$

$$\Leftrightarrow r^{(t+1)} = \frac{\mathbb{E}[n_{....}]}{D} \tag{2.18}$$

Across the region of interest, we can calculate the expected number of transitions between every possible ancestral state by noting that the joint probability of being in ancestral state (u, v) at window w and state (i, j) at window $w + 1$ is

$$P\left(S_w = u, Y_w = v, S_{w+1} = i, Y_{w+1} = j|\mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}\right)$$

$$= \frac{P\left(S_w = u, Y_w = v, S_{w+1} = i, Y_{w+1} = j, \mathbf{X}|\mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}\right)}{P\left(\mathbf{X}|\mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}\right)}$$

$$= \frac{\alpha_w^{(u,v)} \cdot \tau_{w_{ui}} \cdot p_{w+1_j} \cdot m_{i,j} \cdot \beta_{w+1}^{(i,j)}}{P\left(\mathbf{X}|\mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}\right)} \tag{2.19}$$

In this expression, the forward probabilities $\alpha_w^{(u,v)}$ are as defined in eq.(2.12) and the backward probabilities, denoted $\beta_w^{(i,j)}$, of each ancestral state at every locus are

$$P\left(\mathbf{x}_{w+1}, \ldots, \mathbf{x}_W | S_w = i, Y_w = j, \mathbf{q}, r, \mathbf{M}, \mathbf{H}\right)$$

That is, the backward probability is the probability of the observed haplotypes/genotypes from windows $w+1$ to the final window, given that the ancestral state of window w is (i, j). Knowing these probabilities, we can find the expected number of transitions between any two ancestral states, over all windows. It follows that the expected number of total changes in ancestry is

$$\mathbb{E}[n_{....}] = \sum_{u,v,i,j \in \{1,\ldots,K\}} \sum_{w=1}^{W-1} \eta_w^{(u,i)} \cdot P\left(S_w = u, Y_w = v, S_{w+1} = i, Y_{w+1} = j | \mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}\right) \tag{2.20}$$

where

$$\eta_w^{(u,i)} = \frac{(1 - e^{-d_w r^{(t)}}) \cdot q_i^{(t)}}{\mathbb{1}[u = i] \cdot e^{-d_w r^{(t)}} + (1 - e^{-d_w r^{(t)}}) \cdot q_i^{(t)}} \tag{2.21}$$

is the probability that, given a transition occurs between windows w and $w+1$ from underlying ancestral states u to i, recombination took place. If there is a change in the underlying ancestry (and so $u \neq i$) then this probability is equal to one since recombination must have occurred if the underlying ancestry changes between the two windows. If the underlying ancestry is unchanged ($u = i$) then the probability that recombination has occurred but has not resulted in an ancestral switch is $\eta_w^{(u,i)} < 1$. Equations (2.18) to (2.21) now give the calculation of the update of the parameter r in the EM algorithm.

Next we derive the EM-update of the misfitting probabilities by maximizing the

log-likelihood in eq.(2.17) with respect to m_{ij} subject to the constraint that $\sum_{j=1}^{K} m_{ij} = 1$ for all $i \in \{1, \ldots, K\}$. Doing this, the equation to solve becomes

$$\frac{\partial \mathbb{E}[\ell(\mathbf{q}, r, \mathbf{M}|\mathbf{X}, \mathbf{S}, \mathbf{Y})]}{\partial m_{ij}} + \psi = \frac{\sum_{w=1}^{W} \mathbb{E}[\mathbb{1}[(S_w, Y_w) = (i, j)]]}{m_{ij}} + \psi = 0$$

$$\Leftrightarrow m_{ij} = \frac{\sum_{w=1}^{W} \mathbb{E}[\mathbb{1}[(S_w, Y_w) = (i, j)]]}{-\psi} \tag{2.22}$$

where ψ is the Lagrange multiplier. We solve for ψ:

$$\Leftrightarrow \sum_{j=1}^{K} m_{ij} = \frac{\sum_{w=1}^{W} \sum_{j=1}^{K} \mathbb{E}[\mathbb{1}[(S_w, Y_w) = (i, j)]]}{-\psi}$$

$$\Leftrightarrow 1 = \frac{\sum_{w=1}^{W} \mathbb{E}[\mathbb{1}[S_w = i]]}{-\psi}$$

$$\Leftrightarrow \psi = -\sum_{w=1}^{W} \mathbb{E}[\mathbb{1}[S_w = i]]$$

Substituting for ψ in eq.(2.22) gives the update $m_{ij}^{(t+1)}$ to be

$$m_{ij}^{(t+1)} = \frac{\sum_{w=1}^{W} \mathbb{E}[\mathbb{1}[(S_w, Y_w) = (i, j)]]}{\sum_{w=1}^{W} \mathbb{E}[\mathbb{1}[S_w = i]]}$$

To calculate this update, we note that the expectation $\mathbb{E}[\mathbb{1}[(S_w, Y_w) = (i, j)]]$ is equal to the marginal probability of ancestral state $(S_w = i, Y_w = j)$ at window w for the parameter estimates at iteration t. This expectation may be calculated from the forward and backward probabilities:

$$\mathbb{E}[\mathbb{1}[(S_w, Y_w) = (i, j)]]$$

$$= P\left(S_w = i, Y_w = j | \mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}, \mathbf{H}\right)$$

$$= \frac{P\left(\mathbf{x}_1, \ldots, \mathbf{x}_w, S_w = i, Y_w = j | \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}, \mathbf{H}\right) \cdot P\left(\mathbf{x}_{w+1}, \ldots, \mathbf{x}_W | S_w = i, Y_w = j, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}, \mathbf{H}\right)}{P(\mathbf{X} | \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}, \mathbf{H})}$$

$$= \frac{\alpha_w^{(i,j)} \cdot \beta_w^{(i,j)}}{\sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_w^{(i,j)} \cdot \beta_w^{(i,j)}}$$

Lastly, we derive the update of the parameter of global ancestry proportions $\mathbf{q}$. We take derivative of the expected log-likelihood with respect to q_i and maximize it under the constraint that $\sum_{i=1}^{K} q_i = 1$, the equation to solve is

$$\frac{\partial \mathbb{E}[\ell(\mathbf{q}, r, \mathbf{M} | \mathbf{X}, \mathbf{S}, \mathbf{Y})]}{\partial q_i} + \psi = \frac{\partial \mathbb{E}[\mathbb{1}[S_1 = k] log(q_i) + n_{..i.} log(q_i)]}{\partial q_i} + \psi = 0$$

$$\Leftrightarrow \frac{\mathbb{E}[\mathbb{1}[S_1 = k]] + \mathbb{E}[n_{..i.}]}{q_i} + \psi = 0$$

$$\Leftrightarrow q_i = \frac{\mathbb{E}[\mathbb{1}[S_1 = k]] + \mathbb{E}[n_{..i.}]}{-\psi} \tag{2.23}$$

where ψ is the Lagrange multiplier which we find by taking the sum over all populations:

$$\sum_{i=1}^{K} q_i = \frac{\sum_{i=1}^{K} \left(\mathbb{E}[\mathbb{1}[S_1 = k]] + \mathbb{E}[n_{..i.}]\right)}{-\psi}$$

$$\Leftrightarrow 1 = \frac{1 + \mathbb{E}[n_{....}]}{-\psi}$$

$$\Leftrightarrow \psi = -(1 + \mathbb{E}[n_{....}])$$

It follows from eq.(2.23) that the EM update for $\mathbf{q}$ is

$$q_i^{(t+1)} = \frac{\mathbb{E}[\mathbb{1}[S_1 = k]] + \mathbb{E}[n_{..i.}]}{1 + \mathbb{E}[n_{....}]} \tag{2.24}$$

where the expectation is found for the current value of the parameter set. The first

expectation $\mathbb{E}[\mathbb{1}[S_1 = k]]$ is simply the marginal probability of the underlying ancestry being population k at the first window. Recall that the second expectation $\mathbb{E}[n_{..i.}]$ is the expected number of recombination events that result in the new underlying ancestry being of population i. That is, it includes the occurrence of recombination events that do not result in a change in the underlying ancestry state. As is similar to eq.(2.20) but this time summing over all indices except i, we have that

$$\mathbb{E}[n_{..i.}] = \sum_{u,v,j \in \{1,\ldots,K\}} \sum_{w=1}^{W-1} \eta_w^{(u,i)} \cdot P\left(S_w = u, Y_w = v, S_{w+1} = i, Y_{w+1} = j | \mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}\right) \tag{2.25}$$

where $\eta_w^{(u,i)}$ is as defined in eq.(2.21).

For these updated parameter values $(\mathbf{q}^{(t+1)}, r^{(t+1)}, \mathbf{M}^{(t+1)})$, we once again compute the forward and backward probabilities and repeat the parameter updates until the algorithm converges. In practice, the algorithm may be considered to have converged when the absolute change in the parameters at an update is smaller than some prespecified value. The EM algorithm will converge to a local maximum and not necessarily the global maximum so several runs starting at a range of initial parameter values is recommended to ensure that a global maximum is reached. In practice, convergence was found to be extremely fast so it is still practical for many runs to be performed. As in the MCMC approach, we classify the inferred underlying ancestral state $\hat{z_w}$ to be that which has the highest marginal posterior probability, these marginal probabilities being calculated from the forward and backward probabilities once the algorithm has converged.

2.4.4 CEM algorithm

The Classification EM (CEM) algorithm is a variation on the EM approach just described that makes use of the Viterbi algorithm to find the sequence of ancestral states

that are *jointly* most likely for a particular set of parameter values given the observed data (Celeux and Govaert, 1992). This sequence is known as the Viterbi path. Whereas in the EM algorithm we find the expectation of the complete data log-likelihood over the hidden states, in the CEM we estimate the hidden states given the observed data. In a sense, the EM method is the more fastidious approach of the two, as it takes in to account all possibilities of the hidden ancestral states and finds the local maximum of the expected complete-data likelihood over the set of all possible unobservable paths. The CEM on the other hand, directly estimates the hidden states at every iteration for the current parameter estimates, and then treats them as though they are observed data from which the model parameter estimates are then updated to maximize the likelihood of these hidden states.

The CEM is initialised at some parameter values for which the Viterbi path is found. Taking this sequence to be the current estimate of the hidden data, the parameters are updated to maximize the likelihood of the Viterbi path. The path itself may be viewed as a parameter in the model in the CEM approach. For these new parameter values the Viterbi sequence is obtained once again and the parameter updates are repeated. This iterative process continues until convergence. Written explicitly in the case of our model, the CEM algorithm consists of two steps:

- Classification : we find the Viterbi path $\{\gamma_1^{(t)}, \ldots, \gamma_W^{(t)}\}$ such that

$$\{\gamma_1^{(t)}, \ldots, \gamma_W^{(t)}\} = \underset{\gamma_1,\ldots,\gamma_W}{\arg\max}\, P(\gamma_1, \ldots, \gamma_W | \mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)})$$

  where the maximization is found over the set of all possible paths and the state at each window γ_w consists of two components (s_w, y_w) denoting the underlying and fitted ancestries respectively.

- Maximization : given the Viterbi path found in the above step, we update the

parameter set such that

$$\{\mathbf{q}^{(t+1)}, r^{(t+1)}, \mathbf{M}^{(t+1)}\} = \arg\max_{\mathbf{q},r,\mathbf{M}} P(\gamma_1^{(t)}, \ldots, \gamma_W^{(t)} | \mathbf{q}, r, \mathbf{M})$$

The structure of our HMM allows a straight forward calculation of the Viterbi path through a dynamic programming approach known as the Viterbi algorithm. At the tth iteration, for each window in turn, we calculate the probability $v_{ij}(w)$ of the most likely path that ends in state $(S_w = i, Y_w = j)$ at window w, that is

$$v_{ij}(w) = \max_{S_1,Y_1,\ldots,S_{w-1},Y_{w-1}} P(\mathbf{x}_1, \ldots, \mathbf{x}_w, S_1, Y_1, \ldots, S_{w-1}, Y_{w-1}, S_w = i, Y_w = j | \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}) \tag{2.26}$$

where $(\mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)})$ are the parameter estimates at iteration t. The algorithm is initialized at the first window where, for $(S_1 = i, Y_1 = j)$

$$v_{ij}(1) = q_i^{(t)} \cdot m_{ij} \cdot p_{1_j}$$

The recurrence relation that describes how this probability is calculated progressively along the windows is

$$v_{ij}(w+1) = p_{w+1_j} \cdot m_{ij} \cdot \max_{u,v \in \{1,\ldots,K\}} [v_{u,v}(w) \cdot \tau_{w_{ui}}]$$

The key is to record, for each state at every window $w+1$, which state $\mu_w(ij)$ at the previous window maximized the probability in eq.(2.26). Explicitly,

$$\mu_w(ij) = \arg\max_{u,v \in \{1,\ldots,K\}} [v_{uv}(w-1) \cdot \tau_{w-1_{ui}}]$$

At the last window, the probability P_v of the Viterbi path is

$$P_v = \max_{i,j \in \{1,\dots,K\}} v_{ij}(W)$$

and the state of the Viterbi path at this window is

$$\gamma_W = \arg\max_{i,j \in \{1,\dots,K\}} v_{ij}(W)$$

Knowing this final state, we can recall the state at the penultimate window that maximizes eq.(2.26), and so on for window $W-2,\dots,1$. The Viterbi path $\{\gamma_1,\dots,\gamma_W\}$ is obtained by this backward recall relation:

$$\gamma_w = \mu_{w+1}(\gamma_{w+1})$$

Once the Viterbi path has been found, the next step is to update the parameter values such that they maximize the likelihood of this path. As in the EM algorithm, the update of r is that of eq.(2.18) however in the CEM method the expectations are taken over the Viterbi path of hidden states found in the previous step, rather than over all possible hidden paths as in the EM. That is,

$$\mathbb{E}[n_{....}] = \sum_{u,v,i,j \in \{1,\dots,K\}} \sum_{w=1}^{W-1} \eta_w^{(u,i)} \cdot \mathbb{1}\left[\gamma_w = (u,v), \gamma_{w+1} = (i,j) | \mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}\right] \quad (2.27)$$

where u is the underlying population and v is the fitted population of the Viterbi state γ_w at window w. Similarly, i is the underlying population and j is the fitted population of the Viterbi state γ_{w+1} at window $w+1$. The update of the global ancestry proportion q_i is that of eq.(2.24) except that, as in the update of r, the expectation is taken over the

Viterbi path so

$$\mathbb{E}[n_{..i.}] = \sum_{u,v,j \in \{1,\ldots,K\}} \sum_{w=1}^{W-1} \eta_w^{(u,i)} \cdot \mathbb{1}\left[\gamma_w = (u,v), \gamma_{w+1} = (i,j) | \mathbf{X}, \mathbf{q}^{(t)}, r^{(t)}, \mathbf{M}^{(t)}\right] \tag{2.28}$$

The CEM estimates of the misfitting probabilities are updated as follows:

$$m_{ij}^{(t+1)} = \frac{\sum_{w=1}^{W} \mathbb{1}[s_w = i, y_w = j]}{\sum_{w=1}^{W} \mathbb{1}[s_w = i]} \tag{2.29}$$

where (s_w, y_w) is the state of the Viterbi path at window w. This equates to simply finding, among the states that the make up the Viterbi path, of those where the underlying ancestry is population i the proportion for which the fitted ancestry is population j. This two step procedure of finding the Viterbi algorithm and updating the parameters $\mathbf{q}$, r and $\mathbf{M}$ is repeated until convergence. Once the CEM algorithm has converged, the local ancestry calls are those of the Viterbi path in the final iteration.

This concludes the exposition of the three different methods that we have used to implement the MULTIMIX model - the MCMC, EM and CEM algorithms. Table 2.1 summarises the settings in which each method may be applied, which parameters are estimated and which are set by the user. We note that the extension of the EM and CEM methods to handle unphased admixed samples is an obvious further extension of MULTIMIX that we would like to implement in future versions. Furthermore, it would be straight forward to include the estimation of the misfitting probabilities in the MCMC algorithms and this may also be incorporated in future work.

MULTIMIX Method	Panel phasing	Sample phasing	Estimated parameters	User-specified parameters
MCMC	either phased or unphased	either phased or unphased	$\mathbf{q}$, r	n, λ, $\mathbf{M}$
EM	either phased or unphased	phased	$\mathbf{q}$, r, $\mathbf{M}$	n, λ
CEM	either phased or unphased	phased	$\mathbf{q}$, r, $\mathbf{M}$	n, λ

Table 2.1: Summary of settings in which the different MULTIMIX methods may be applied.

2.5 A Model of Conditional Misfitting

So far the model of misfitting that we have described considers that this process happens independently at each window. That is, the probability that misfitting occurs at a window is independent of whether or not misfitting took place at the preceding window. This is a simple model that has been shown through simulations to perform well, however we have observed that in some instances the model misses very short chunks of ancestry, only a couple of windows in length. The motivation for introducing a conditional miscopying model is to try to improve performance at these very narrow windows.

The modification is that the probability of fitting to population j at window w when the underlying ancestry is population i is now also dependent on the fitted and underlying states of the preceding window. That is, rather than the expression in eq.(2.8), this conditional misfitting probabilities are now

$$P(Y_w = j | S_w = i, Y_{w-1} = v, S_{w-1} = u) = m_{ijuv} \qquad i, j, u, v \in \{1, \ldots, K\}.$$

These probabilities may be expressed as

$$P(Y_w = j|S_w = i, Y_{w-1} = v, S_{w-1} = u, \mathbf{X})$$

$$= \frac{P(Y_w = j, S_w = i, Y_{w-1} = v, S_{w-1} = u, \mathbf{X})}{\sum_{j=1}^{K} P(Y_w = j, S_w = i, Y_{w-1} = v, S_{w-1} = u, \mathbf{X})}$$

where

$$P(Y_w = j, S_w = i, Y_{w-1} = v, S_{w-1} = u, \mathbf{X}) = \alpha_{w-1}^{(u,v)} \cdot \tau_{w-1_{ui}} \cdot p_{w_j} \cdot m_{ijuv} \cdot \beta_w^{(ij)}$$

This approach equates to estimating separate misfitting matrices for each possible ancestry state (S, Y) of the preceding window.

2.6 Resolving boundaries

Due to the use of windows of SNPs in our method, there will be some unavoidable misclassification of ancestry in windows that encompass sites where the ancestry changes. The number of SNPs incorrectly classified in this manner will depend on the extent of admixture and the window size. To deal with this, we require some post-processing of the local ancestry estimates in order to determine more accurately the site at which the switch occurs.

To begin resolving the boundaries at switch points, we take each pair of consecutive windows in turn for which the estimated ancestry is different and consider that each locus in these two windows is a candidate for the site of the switch. We focus on an extended region that includes this pair along with some number of flanking windows, n_f, either side to give a total of N loci in the region where $N = (2n_f + 2)w$. Let k_1 and k_2 be the ancestry that has been called at the first and second window of the pair

respectively. The boundary will be tested at all positions j such that it defines a chunk to the left of it containing loci $\{1, 2, \ldots, j\}$ assumed to be of ancestry from population k_1, and another to its right composed of loci $\{j+1, \ldots, N\}$ for $j \in \{n_f w + 1, \ldots, (n_f + 2)w - 2\}$ assumed to be of ancestry k_2.

For each position j, we calculate the sum of the multivariate normal log-densities f_l and f_r of the sample at the left chunk $\mathbf{x}_l$ and at the right chunk $\mathbf{x}_r$, given their assumed ancestries y_l and y_r. That is we find $\ell(j)$:

$$\ell(j) = log f_l(\mathbf{x}_l | y_l = k_1) + log f_r(\mathbf{x}_r | y_r = k_2)$$

where $\mathbf{x}_l = \{x_1, \ldots, x_j\}$ and $\mathbf{x}_r = \{x_{j+1}, \ldots, x_N\}$. The parameters of these two distributions are estimated from the corresponding ancestral panels, as described in section 2.3.1. We then estimate that the boundary lies at position J where $J := arg_j \; max \; \ell(j)$ with the view that if the boundary is indeed at the true position the density of the sample across the left and right chunks will be greater than when it is placed elsewhere in the region.

If it happens that there are two inferred boundaries within a single window length of each other, then it would be naive to resolve each side of the chunk independently across the whole length of the window. In this case the positions of the two resolved boundaries may conflict each other as the ancestry at some SNPs within the window may be classified differently by the two resolving steps. Instead, the set of positions over which each boundary is tested may be restricted to only half of those in the window, preventing any over lap of the boundaries at either side of the chunk.

This alteration of the boundary resolving algorithm is not currently implemented in our MULTIMIX method, but it is an extension that may be added in future work to develop the model. In realistic levels of admixture, chunks of ancestry less than a couple of hundred SNPs in length are extremely rare and we do not expect them to

have a considerable effect on the performance of our method. We will show in Chapter 3, via simulation studies, that even without this extension the current approach to resolving boundaries works well.

2.7 Summary

In this chapter we described two methods of ancestry estimation that are applicable to studying an admixed individual who has been genotyped, and possibly phased, at a dense set of linked SNPs for which there are also genotyped samples available from candidate ancestral populations. Both methods may be used to analyse a single study individual and do not require multiple admixed individuals to perform inference.

We started by introducing a fast and accurate method that uses the EM algorithm to estimate global ancestral proportions using a filtered set of SNPs. The model underlying this approach is a simplified version of that used in the STRUCTURE, FRAPPE and ADMIXTURE methods, and while it is useful in estimating global ancestry it does not describe local ancestry.

The second and more sophisticated method we presented does perform local inference. We assume a multivariate normal distribution on the observed study haplotype within contiguous windows along a chromosome, conditional upon the ancestral state of the window. The parameters of the distribution for each population are estimated from the corresponding ancestral panel. Changes in ancestry across windows are modeled as a Poisson process whose rate is a function of the number of generations since admixture which is estimated in the model fitting. In any window, we allow the population to which the MVN distribution is fit to differ from that of the true underlying population - a characteristic of the model that we refer to as *misfitting*. We show how three different statistical techniques (MCMC sampling, an EM algorithm and a CEM algorithm) may be used to fit the model. Finally, we presented a post-processing step

to resolve the location of boundaries between sites of ancestral switches.

In the next chapter we carry out a comprehensive investigation of the performance of these three techniques via simulation. We investigate the effect of the parameters that are fixed in the model and demonstrate that the performance of the model in a two-way admixture scenario is comparable to that of the leading method HAPMIX. We look at what effect phasing of the panel and of the study individual have on model performance and conduct a three-way admixture simulation to test the model in a more challenging setting.

Chapter 3

Simulations to Assess Model Performance

In Chapter 2 we presented a novel model to describe local ancestry in an admixed individual and described three statistical techniques that may be used in model fitting. In this chapter we conduct a simulation study to investigate the performance of our method in two-way and multi-way admixture scenarios. We assess the effect of the window size, λ and misfitting parameters. We show that we achieve an accuracy comparable to that of HAPMIX in the two-way setting, and we look at how phasing of the ancestral panels and the study individual affects model performance.

3.1 Simulating admixed individuals

In order to assess the performance of our model we conducted a simulation study in which we generated many admixed genomes and tested how well our method deduces their local ancestry. We simulated these admixed individuals under various settings of ancestral proportions and time since admixture to provide a range of extent of admixture, using the samples from the following populations of HapMap Phase 3 :

Utah residents with ancestry from northern and western Europe (CEU); Han Chinese from Beijing (CHB); Yoruba from Ibadan, Nigeria (YRI) and the Gujarati Indians (GIH) living in Houstan, Texas. We also make use of the Native American samples that are present in the HGDP data set. These consist of 28 Pima, 42 Maya, 14 Colombian, 26 Karitiana and 16 Surui haplotypes. We investigate how well it does in different admixture scenarios involving various numbers of these populations, where in all cases the admixed genomes are simulated according to the mode described below.

The ancestry along one copy of a chromosome is simulated according to the following switching model. The ancestry of the first SNP Z_1 is chosen randomly with probabilities according to the specified global ancestry proportions for each of the two populations in the admixture. Given this ancestral state k at SNP i, the population of origin $k^{'}$ of the next SNP in the sequence is modeled as a Markov process. They are sampled according to the following transition probabilities where d_i is the genetic map distance in Morgans between loci i and $i+1$, r is the switching rate parameter and $\mathbf{q}$ is the vector of global ancestry proportions:

$$P(z_{i+1} = k^{'} | z_i = k, r, \mathbf{q}) = \begin{cases} P(\text{no switch or some amount of switching resulting} \\ \text{in the same ancestral state } k) & \text{if } k^{'} = k \\ P(\text{some amount of switching resulting} \\ \text{in the ancestral state } k^{'}) & \text{otherwise} \end{cases}$$

$$= \begin{cases} e^{-d_i r} + (1 - e^{-d_i r}) q_{k'} & \text{if } k^{'} = k \\ (1 - e^{-d_i r}) q_{k'} & \text{otherwise} \end{cases}$$

Next, conditional upon these ancestral states we then simulate the corresponding haplotype. For every population we have a set of sampled haplotypes to which we apply the copying model of Li and Stephens (2003). Allowing X_i to denote the index number of the haplotype being copied at SNP i, X_i is modeled as a Markov chain,

initialized with $P(X_1 = x) = \frac{1}{k}$ where k is the number of sampled haplotypes available from that population. The transition probabilities are:

$$P(X_{i+1} = x^{'}|X_i = x) = \ = \begin{cases} e^{\frac{-\rho_i d_i}{k}} + (1 - e^{\frac{-\rho_i d_i}{k}}) * \frac{1}{k} & \text{if } x^{'} = x \\ (1 - e^{\frac{-\rho_i d_i}{k}}) * \frac{1}{k} & \text{otherwise} \end{cases}$$

where d_i is the physical distance between SNPs i and $i + 1$ and $\rho_i = 4Nc_i$ where N is the effective diploid population size and c_i is the average rate of cross over per unit physical distance, per meiosis, between the two consecutive SNPs. It follows that $c_i d_i$ is the genetic distance between the SNPs i and $i + 1$.

To include the occurrence of mutations, we allow the copying process to be imperfect with probability $\frac{\tilde{\theta}}{k+\tilde{\theta}}$, where $\tilde{\theta}$ is a per site mutation parameter . That is, if h_{ji} denotes the allele of haplotype j at locus i then

$$Pr(h_{k+1i} = a|X_i = x, h_1, \ldots, h_k) = \begin{cases} \frac{k}{k+\tilde{\theta}} + \frac{1}{2} \cdot \frac{\tilde{\theta}}{k+\tilde{\theta}} & h_{xi} = h_{X_i i} \text{ i.e. perfect copying} \\ \frac{1}{2} \cdot \frac{\tilde{\theta}}{k+\tilde{\theta}} & h_{xi} \neq h_{X_i i} \text{ i.e. imperfect copying} \end{cases}$$

At a single site on a genealogical tree relating n random chromosomes, the expected number of mutation events is $\tilde{\theta} \sum_{m=1}^{n-1} \frac{1}{m}$, so setting $\tilde{\theta} = (\sum_{m=1}^{n-1} \frac{1}{m})^{-1}$ gives *a priori* the expected number of mutation events at each locus to be 1.

In this manner, we generate each admixed individual independently from a panel of ancestral haplotypes at a set of linked SNPs. We specify the number of generations since admixture and the global proportions of ancestry to control the extent of admixture.

3.2 African-American Simulations

In order to gauge the performance of our method in a realistic admixture scenario, we conducted a study of individuals simulated to be representative of African Americans -

a population that is often investigated by ancestry estimation methods. We compared how well MULTIMIX estimated local ancestry in comparison to HAPMIX, the current leading method. The ancestral panels of haplotypes came from HapMap. These samples that we used consisted of 234 CEU and 230 YRI haploid autosomes from unrelated individuals however only half of these were used in the simulations while the other half were used in model testing. Effective population sizes of 11400 and 17400 were used for the CEU and YRI respectively. We set q, the global proportion of CEU ancestry, to be 0.2 and $r \in \{5, 10, 50\}$ where r is the rate of recombination events per Morgan in the admixed chromosomes, only some of which result in ancestry switches. For each value of r we simulated 10 individuals by generating each copy of their autosomes separately, for a total of 30 individuals simulated of CEU-YRI admixture, at approximately 1.2 million SNPs genome-wide. These simulated individuals were analyzed using each of the three implementations of the MULTIMIX model (1) MCMC (2) EM and (3) CEM. For each of these methods, performance is then measured as the proportion of SNPs at which this estimated underlying ancestry is equal to the true ancestry. We report the performance both before and after the step to resolve boundaries has been conducted.

3.2.1 Estimating global ancestry proportions

The model described in section 2.2 allows us to estimate the global ancestral proportions $\mathbf{q}$ of the admixed samples using a subset of SNPs that are in approximate linkage equilibrium with each other. Such a subset was obtained using Plink (Purcell et al., 2007) to prune within windows of 50 SNPs at a time. Plink recursively removes SNPs if their multiple correlation coefficient R^2 when simultaneously regressed on all the others in that window is greater than some specified value. The position of the window is then shifted by 5 SNPs and the pruning process is repeated at the new location.

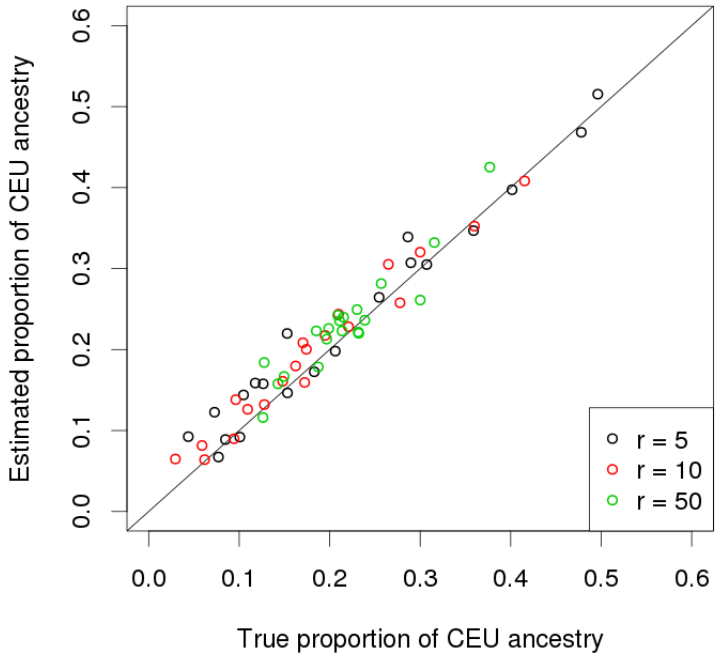

Figure 3.1: The estimates of the global CEU ancestry proportions of the CEU-YRI simulated individuals as estimated by the model on unlinked SNPs.

We pruned the SNP set at $R^2 = 0.33$ and the resulting subset retained 11,587 SNPs of the 100,874 on chromosome 1.

The EM algorithm used to estimate $\mathbf{q}$ is very quick to converge, with 100 iterations being performed on the 60 simulated copies of chromosome 1 in less than a minute where the mode to implement the method was written in C. Figure 3.1 shows that the estimates obtained by the model were accurate for samples simulated at all three values of $r \in \{5, 10, 50\}$. The mean squared error of the estimates was 8.09×10^{-4}, demonstrating that this naive and simple model that does not make use of LD information and that ignores correlations in ancestry is a useful tool for global ancestry estimation.

3.2.2 Investigation of model parameters n, λ and m.

In the MULTIMIX model, the number of SNPs per window n and the term λ in eq.(2.4) are fixed parameters in the model. Both these parameters may influence the performance of the method. A smaller n is preferable if we want to obtain a high-resolution estimate of Z. On the other hand, as n gets smaller the less valid it will be to ignore an increasing dependency between contiguous windows of shared ancestry which may bias inference. Larger windows would be expected to capture more variation between the ancestral populations, but they are also more likely to overlap sites of ancestral switches and therefore contains SNPs of different ancestry. The optimal value of the window size will best balance this trade-off.

A non-zero λ is required to ensure the covariance matrix of the MVN model is invertible so that the calculations can be efficiently carried out in each window. Increasing λ will tend to reduce the ability of the model to discriminate the ancestry of a given haplotype within each window. When developing our algorithm using simulated data of known ancestry we often found that within individual windows the MVN model would sometimes clearly prefer to infer an ancestry other than the true ancestry. Increasing λ may help to smooth out these errors, as will our use of both fitted and true ancestral states, $\mathbf{Y}$ and $\mathbf{S}$, respectively.

We analysed all 30 simulated individuals using varying levels of n and λ. We considered all combinations of $n \in \{50, 100, 150\}$ and $\lambda \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$ in the analysis using the MCMC algorithm. For this analysis we fixed the misfitting probabilities at $m_{12} = m_{21} = b$ for $b \in \{0.05, 0.1\}$. Both the simulated samples and the ancestral panels were phased in this analysis. Performance was measured as the number of loci at which the method correctly ascertained the underlying ancestral population across all 10 samples for each value of r. An example of how the estimated local ancestry as inferred by MCMC-MULTIMIX compared to the true simulated ancestry along a single

copy of chromosome 1 is illustrated in Fig.3.2. The first two plots show the simulated and estimated ancestry of a single copy of chromosome 1 for 10 generations of mixing, while the latter two plots show an example for much older admixture of 50 generations of mixing. In both of these examples the window size used was 100 SNPs, λ was set to be 0.005, the misfitting probabilities m_{12} and m_{21} were 0.05 and the step to resolve the precise location of the boundaries between the sites of switches has been applied. We can see that in both of these examples there is excellent agreement between the ancestry calls made by MULTIMIX and the true ancestry, even in the case of 50 generations of mixing where our method is able to detect the many narrow chunks of CEU ancestry among a YRI background, however not perfectly as the figure shows that one such chunk is missed altogether in each of these examples.

Having run the MCMC implementation of MULTIMIX on the three sets of CEU-YRI individuals, simulated with different values of r, over a range of parameter values we found that 100 SNPs was the optimal window size in the analysis of each group (Table 3.1). A bootstrap estimate of the standard error of the performance was found to be 0.104% for simulations in the $r = 10$ parameter group meaning that the highest performance value is not significantly different from that in some of the other parameter settings. Nevertheless, this analysis serves to identify which parameter values to set as default when using MULTIMIX. For the $r = 10$ and $r = 50$ groups, a value of $\lambda = 0.005$ was best, while for $r = 5$ setting $\lambda = 0.01$ gave the most correct ancestry calls. Setting the misfitting probabilities m_{12} and m_{21} to be 0.05 was optimal for $r = 5$ and $r = 50$ and a slightly higher value of 0.1 was best for the $r = 10$ group. Over all, when considering the accuracy of ancestry calls for all 30 simulated individuals, the optimal parameter settings were $n = 100$, $\lambda = 0.005$ and $m_{12} = m_{21} = 0.05$ and we would recommend these values when using the MCMC-MULTIMIX.

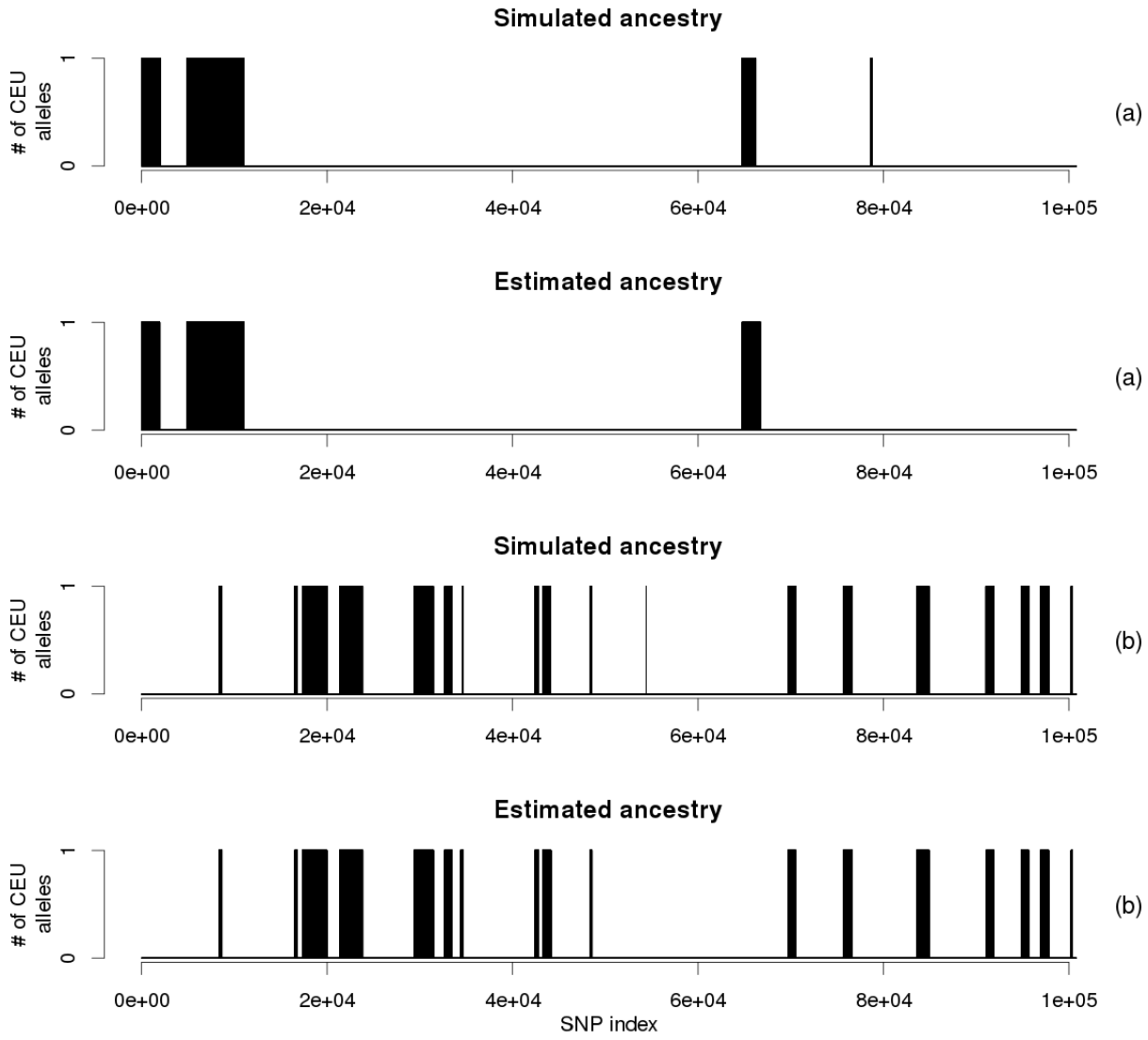

Figure 3.2: A comparison of the local ancestry calls made by MCMC-MULTIMIX and the true simulated ancestry along a single copy of chromosome 1 simulated to be 20% CEU and 80% YRI for (*a*) 10 generations of mixing and for (*b*) another copy for 50 generations.

n	λ	m	$r = 5$	resolved	$r = 10$	resolved	$r = 50$	resolved	Over all	resolved
50	0.005	0.05	98.133	98.117	98.033	98.035	95.994	96.095	97.386	97.416
50	0.005	0.1	99.203	99.244	99.097	99.144	97.248	97.436	98.516	98.608
50	0.01	0.05	98.148	98.126	98.004	97.991	96.098	96.152	97.417	97.423
50	0.01	0.1	99.316	99.342	99.030	99.073	97.165	97.297	98.504	98.571
50	0.05	0.05	97.787	97.763	97.590	97.588	95.732	95.763	97.036	97.038
50	0.05	0.1	99.153	99.163	98.718	98.763	96.734	96.837	98.202	98.254
50	0.1	0.05	97.657	97.627	97.448	97.465	95.454	95.542	96.853	96.878
50	0.1	0.1	99.055	99.071	98.636	98.657	96.462	96.624	98.051	98.117
50	0.5	0.05	98.796	98.837	98.344	98.415	95.560	95.864	97.560	97.705
50	0.5	0.1	99.302	99.336	98.845	98.908	96.234	96.534	98.127	98.259
100	0.005	0.05	99.728	99.794	**99.540**	99.640	**97.701**	**98.242**	**98.990**	**99.225**
100	0.005	0.1	99.723	99.793	99.531	**99.644**	97.549	98.015	98.934	99.151
100	0.01	0.05	**99.738**	**99.797**	99.530	99.641	97.691	98.142	98.987	99.193
100	0.01	0.1	99.731	99.796	99.539	99.631	97.582	98.012	98.950	99.146
100	0.05	0.05	99.568	99.610	99.477	99.579	97.427	97.862	98.824	99.017
100	0.05	0.1	99.686	99.733	99.509	99.599	97.531	97.938	98.909	99.090
100	0.1	0.05	99.537	99.596	99.308	99.400	97.142	97.587	98.662	98.861
100	0.1	0.1	99.670	99.734	99.484	99.564	97.405	97.808	98.853	99.036
100	0.5	0.05	99.425	99.519	99.116	99.246	96.446	96.991	98.329	98.585
100	0.5	0.1	99.532	99.628	99.134	99.245	96.565	97.067	98.410	98.647
150	0.005	0.05	99.628	99.724	99.380	99.541	97.290	98.060	98.766	99.108
150	0.005	0.1	99.637	99.738	99.363	99.496	96.839	97.483	98.613	98.906
150	0.01	0.05	99.674	99.754	99.433	99.589	97.334	98.031	98.813	99.125
150	0.01	0.1	99.675	99.761	99.367	99.507	96.925	97.557	98.656	98.942
150	0.05	0.05	99.624	99.716	99.392	99.553	97.357	98.006	98.791	99.091
150	0.05	0.1	99.656	99.757	99.376	99.517	97.050	97.636	98.694	98.970
150	0.1	0.05	99.620	99.717	99.380	99.527	97.157	97.808	98.719	99.017
150	0.1	0.1	99.658	99.756	99.375	99.502	97.026	97.642	98.686	98.967
150	0.5	0.05	99.493	99.594	99.197	99.378	96.380	97.189	98.357	98.721
150	0.5	0.1	99.559	99.654	99.226	99.414	96.567	97.326	98.451	98.798

Table 3.1: Investigation of the performance of MCMC-MULTIMIX on the individuals simulated to be 20% CEU and 80% YRI for $r \in \{5, 10, 50\}$, for different values of the parameters n, λ and m both before and after resolving the boundaries of switches. The maximum values in each column appear in bold.

Method	$r = 5$	$r = 10$	$r = 50$
HAPMIX	99.915	99.824	99.095
MULTIMIX	99.797	99.650	98.251
Concordance	99.770	99.658	98.345

Table 3.2: Performance of HAPMIX and MULTIMIX on the CEU-YRI simulated individuals, and the concordance between their local ancestry calls. The MULTIMIX performance reported here is the highest value obtained over the MCMC, EM and CEM implementations.

3.2.3 Comparison with HAPMIX

We compared the performance of MULTIMIX with that of HAPMIX as it is also applicable to dense linked SNPs and has been shown to outperform ANCESTRYMAP and LAMP, the latter having been reported to outperform STRUCTURE. We analysed the phased samples of the 30 simulated CEU-YRI individuals with both methods. As HAPMIX requires phased ancestral panels, in this analysis the ancestral data as well as the samples being analysed were phased. We used the haploid mode of HAPMIX with the ANC PROB output setting to produce probabilities of local ancestry, the population for which this is maximal is the classification of the ancestry estimated at each SNP. We set the proportion of CEU ancestry to be 0.2, the mutation values to be 0.2 in CEU, 0.2 in YRI and 0.01 at miscopied sites; the recombination values to be 390 in CEU and 605 in YRI; the miscopying probability to be 0.05; and the number of generations since admixture λ to be $\{6, 11, 51\}$ for the $r = \{5, 10, 50\}$ groups respectively.

We found that HAPMIX slightly outperformed our MULTIMIX method when applied to the CEU-YRI simulated individuals, the results are reported in Table 3.2. However the difference in performance was very small, with the local ancestry calls of the two methods agreeing at 99.26% of SNPS over the three groups simulated at different values of r. Both methods ascertained the local ancestry of the samples less accurately in the case of increased switching.

3.2.4 Comparison of MCMC, EM and CEM methods

We compared the three different methods we have implemented to carry out inference in MULTIMIX. To do this we re-analysed the 30 samples using $n = 100$ and $\lambda = 0.005$ as these were the parameter values found to be optimal over all in the MCMC analysis. For the MCMC method we used fixed values of the misfitting probabilities at $m_{12} = m_{21} = b$ for $b \in \{0.05, 0.1\}$. For the EM and CEM methods we estimated the values of m_{12} and m_{21}. We investigated whether the EM algorithm converges to different local maxima by initializing runs at a wide range of parameter start points consisting of all combinations of $r \in \{2, 10, 50\}$ and $(q_1, q_2) \in \{(0.25, 0.75), (0.5, 0.5), (0.75, 0.25)\}$. When run on all 30 CEU-YRI simulated individuals, initializing the algorithm at 9 different parameter start points and used a window size of 100 SNPs and set $\lambda = 0.005$ for each run, we found that the algorithm converged to the same solution in every run for each sample individual, indicating that the performance of the method is not sensitive to the initial parameter values.

When we compared the percentage of correct ancestry calls made by the MCMC, EM and CEM implementations of MULTIMIX we did not find one method that comprehensively out-performed the others (Table 3.3). In the $r = 5$ group, the MCMC run with $m_{12} = m_{21} = 0.05$ gave the best results, while the EM and CEM performed best in the $r = 10$ and $r = 50$ groups respectively, where in both cases the misfitting probabilities were estimated. All three of the methods offer a useful option of implementing the MULTIMIX model. Each of the three methods were highly accurate in inferring the local ancestry, the main advantage of the EM and CEM version of MULTIMIX being that they deliver estimates extremely quickly and they return an output that is straight forward for the user to process. On the other hand, the MCMC method estimates the posterior distribution of the r and $\mathbf{q}$ parameters, rather than simply returning point estimates as in the EM and CEM, but it is more computationally intense and its output

Method	(m_{12}, m_{21})	$r = 5$	resolved	(m_{12}, m_{21})	$r = 10$	resolved	(m_{12}, m_{21})	$r = 50$	resolved
MCMC	(0.05 , 0.05)	**99.728**	**99.794**	(0.05 , 0.05)	99.540	99.640	(0.05 , 0.05)	97.701	98.242
MCMC	(0.1 , 0.1)	99.723	99.793	(0.1 , 0.1)	99.531	99.644	(0.1 , 0.1)	97.549	98.015
EM	(0.0362 , 0.0386)	99.675	99.736	(0.0414 , 0.0401)	99.546	99.648	(0.0436 , 0.0407)	97.680	98.216
CEM	(0.0352 , 0.0386)	99.683	99.745	(0.0389 , 0.0405)	**99.555**	**99.650**	(0.0360 , 0.0407)	**97.723**	**98.251**

Table 3.3: A comparison of the performance of the three MULTIMIX methods on the CEU-YRI simulated individuals. This misfitting probabilities were fixed in the MCMC analysis but were estimated by the EM and CEM runs. The subscripts 1 and 2 of the misfitting probabilities stand for the CEU and YRI populations respectively. In all of these runs we used a window size of 100 SNPs and $\lambda = 0.005$. The highest percent of correct ancestry calls for each r group is highlighted in bold.

requires more processing than that of the EM and CEM runs.

3.2.5 Inaccurate ancestral panels

The simulations described above tested the performance of both MULTIMIX and HAPMIX under the ideal condition that the ancestral panels contain haplotypes from the same populations from which the admixed individuals were simulated. It is possible that in real applications of these methods that the ancestral populations may not be precisely known, or that there may not be samples available from them.

To test how sensitive the two models are to misspecification of the ancestral haplotypes, we ran MCMC-MULTIMIX (without performing the boundary resolving step) and HAPMIX on all of the CEU-YRI simulated samples but this time used haplotypes from different populations than were used to simulate the admixed samples. In one run, we used the MKK haplotypes of HapMap instead of the YRI as surrogates for the African component of ancestry, and in another run the TSI haplotypes, also of HapMap, were used in place of the CEU panel. We also ran the method with both the African and European panels being misspecified, using both the MKK and the TSI haplotypes at the same time. The MKK and TSI panels consisted of 286 and 176 haplotypes respectively. In MCMC-MULTIMIX the misfitting probabilities were set to 0.05 and a window size of 100 SNPs was used, and λ was set to 0.005. Both the samples and

Method	Panels	$r = 5$	$r = 10$	$r = 50$	Over all
MCMC-MULTIMIX	CEU and YRI	99.728	99.540	97.701	98.990
HAPMIX	CEU and YRI	99.915	99.824	99.095	99.611
MCMC-MULTIMIX	CEU and MKK	99.474	99.221	96.885	98.527
HAPMIX	CEU and MKK	99.420	99.071	97.068	98.520
MCMC-MULTIMIX	TSI and YRI	99.678	99.457	97.492	98.879
HAPMIX	TSI and YRI	99.888	99.804	98.968	99.553
MCMC-MULTIMIX	TSI and MKK	99.480	99.417	97.011	98.636
HAPMIX	TSI and MKK	99.638	99.318	97.508	98.821

Table 3.4: Percent of correct ancestry calls by MCMC-MULTIMIX and HAPMIX when the ancestral panels are not those from which the admixed samples were simulated.

the ancestral panels were phased.

When the CEU and MKK panels were used, MULTIMIX slightly outperformed HAPMIX as it made 98.527% of calls correctly versus HAPMIX at 98.520%. Of the three scenarios tested in Table 3.4, it was this misspecification of the African haplotypes that had the greatest effect on model performance, with a drop in over 1% for HAPMIX and 0.67% for MULTIMIX as compared to when the CEU and YRI panels were used. When the European panel was changed, and the TSI haplotypes were used in place of the CEU, this had the least effect on the performance of both methods as the percent of correct calls by MULTIMIX fell by 0.32% and that of HAPMIX fell by only 0.06% over all. As only around 20% of the ancestry in the admixed samples is European and because the TSI are only very slightly diverged from the CEU and are therefore still an accurate panels of haplotypes, we would expect that replacing the CEU panel with the TSI would have very little effect on the ancestry calls.

3.2.6 Computational performance

We compared the computational performance of the MCMC, EM and CEM implementations of the MULTIMIX model and HAPMIX when analyzing a large set of simulated samples. To do this we generated 400 haploid samples of chromosome 1, simulated to

Method	Time taken
HAPMIX	2.5 hours
MCMC-MULTIMIX	4.5 hours
EM-MULTIMIX	1.1 hours
CEM-MULTIMIX	1.1 hours

Table 3.5: Computational performance of HAPMIX and the MCMC, EM and CEM implementations of MULTIMIX when applied to 400 simulated samples of CEU-YRI admixture along chromosome 1.

be 20%CEU-80%YRI admixed with $r = 10$. For MULTIMIX we used $n = 100$ and $\lambda = 0.005$. The MCMC version was run for a total of 1,000 iterations, the first 400 of which was the burn-in. For the EM and CEM versions we estimated the misfitting probabilities while in the MCMC method they were fixed. All comparisons were carried out on a 3.33GHz Intel Core 2 Duo processor with 3.8GB of RAM.

The time taken by the different methods is reported in Table 3.5. We found that when analysing this large number of chromosomes, the EM and CEM implementation of MULTIMIX were equally fast, taking just over one hour each. The CEM and EM MULTIMIX runs were extremely quick to converge, all samples converging within 7 and 17 iterations respectively. In both of these methods, 90% of the computation time is spent calculating the mean and covariance parameters of the MVN distribution at each window across the chromosome and then finding the log-density of the study samples under these MVN distributions. Nevertheless, these implementations of MULTIMIX were still over twice fast as HAPMIX which took 2.5 hours to analyze the samples. The MCMC method was the most time consuming to run, taking 4.5 hours to perform 1,000 iterations.

3.2.7 Use of unphased data

A key feature of MULTIMIX is that it can handle any combination of phased or unphased study samples and source populations. To determine the effect that phasing

the data has on model performance, we analysed the $r = 10$ group of simulated CEU-YRI individuals for all four combinations of whether the study samples and source populations are either phased or unphased. The MCMC method was used for these comparisons and the performance reported in this section is that obtained without re-solving the boundaries.

As shown in section 3.2.2 when analysing the 20 haploid admixed samples in the $r = 10$ group via the MCMC method with phased CEU and YRI ancestral panels, the highest performance of MCMC-MULTIMIX was 99.540%. In this setting, the optimal parameter values were found to be $\lambda = 0.005$ and the probability of misfitting $m = 0.05$. To assess what effect the absence of phase information for the admixed samples has on the model, we combined the simulated haplotypes in pairs to generate diploid genotypes and ran the method on these unphased samples, once with the ancestral panels consisting of phased haplotypes and again with them being composed of unphased genotypes. These runs, along with the run in which the samples are phased but the panels are unphased, were conducted for a range of parameter values of $\lambda \in \{0.005, 0.01, 0.05, 0.1\}$ and $m \in \{0.05, 0.07, 0.075, 0.08, 0.1\}$. It was found that setting $\lambda = 0.05$ and $m = 0.075$ were the optimal values when analysing the unphased samples for both phased and unphased ancestral panels. In the case of phased samples and unphased panels $\lambda = 0.005$ was the optimal parameter, as was the case when the phased samples were analysed with phased panels, but this time a slightly higher misfitting probability of $m = 0.1$ was optimal. The performance of MCMC-MULTIMIX in each of the four phasing scenarios is reported in Table 3.6. In all four of these scenarios, it is the percent of correct diploid ancestry calls that is reported. That is, whether or not the admixed sample was phased, the corresponding diploid ancestry estimate (either 0, 1 or 2 CEU alleles, say) was compared to the true diploid ancestry. This was done to make the performance on the phased and unphased samples comparable.

In these comparisons, we found that knowing the phase of both the admixed sam-

ples and the ancestral panels was preferable, but that using unphased panels with a phased study sample resulted in only a very small drop in performance of 0.050%. A more considerable decrease of approximately 0.726% in the model's ability to correctly call local ancestry was seen when instead we kept the panels phased but analysed the samples as unphased genotypes. Since HAPMIX is also applicable in the case of phased panels and unphased samples, we ran HAPMIX in this setting and found the performance to be 99.274% which was a drop of 0.34% as compared to a performance of 99.610% found when the phase of the samples was known and diploid calls were made. This shows that the order of the fall in performance when the phase of the samples in unknown (and the panel is phased) is approximately twice as high for MULTIMIX as it is for HAPMIX. When both the samples and the panels were unphased the performance was slightly lower than when the panels were phased but the samples were not, falling by only 0.076%. In Fig.3.3 we illustrate the expected number of CEU alleles at each SNP along chromosome 1 for a diploid individual when analyzed under the four phasing scenarios. While there is very good agreement between these estimates and the true simulated ancestry in each of the scenarios, we can see that in the two cases when the sample is unphased the model misses a very narrow stretch of YRI-YRI ancestry located near SNP number 22,000 among a longer block of a CEU-YRI background that is detected when the sample is phased.

These results show that the MULTIMIX model is able to accurately infer local ancestry for any combination of phased or unphased admixed samples and ancestral panel data. This flexibility is a strength of our method which will be particularly beneficial when accurate phasing is not possible due to a limited number of ancestral samples and lack of related individuals such as trios.

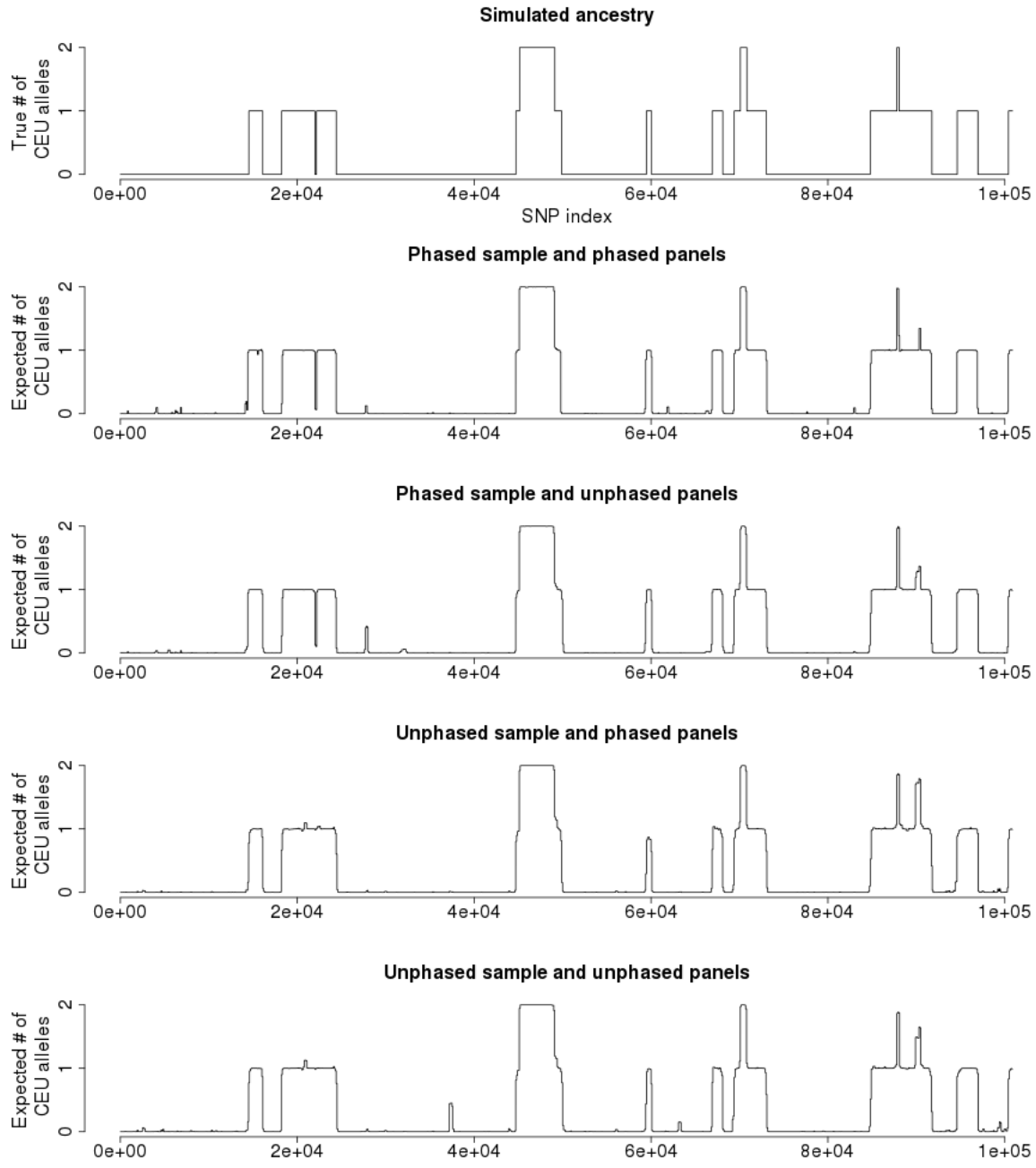

Figure 3.3: Local expectations of the number of CEU alleles at each window for chromosome 1 of an individual simulated to be 20% CEU and 80% YRI with $r = 10$. The true ancestry is plotted in the first figure. The MCMC-MULTIMIX method was applied in four different settings: both the panels and the sample are phased, unphased panels and phased sample, phased panels and unphased sample, and finally both the panels and the sample are unphased.

	phased sample	unphased sample
phased panel	99.144	98.418
unphased panel	99.094	98.342

Table 3.6: A comparison of the performance of MCMC-MULTIMIX in the four phasing scenarios when analysing the CEU-YRI simulated individuals for $r = 10$.

3.2.8 Testing the conditional misfitting model

We wanted to test how the conditional misfitting model described in section 2.5, in which the probability of misfitting at a particular window is conditional on the state the previous window, affects model performance as compared to the original model. This conditional misfitting model was employed in the HMM and the EM algorithm was used to estimate the local ancestry states. In this analysis, both the ancestral panels and the samples were phased. We used a window size of 100 SNPs and set $\lambda = 0.005$. We found that the performance for the $r = 5, 10, 50$ groups was 99.690%, 99.526%, and 97.600% respectively, and 98.938% over all groups collectively. This model fared slightly worse than the original model with the over all performance being 0.0273% lower than that of the unconditional misfitting model.

Recall that in the conditional misfitting model, there are separate misfitting probabilities for each ancestral state of the preceding window. That is, the misfitting probabilities are m_{ijuv} for $i, j, u, v \in \{1, \ldots, K\}$ where i is the underlying ancestral population and j is the fitted population at the current window, and u and v are the underlying and fitted populations at the previous window. In this case, $K = 2$ where the index 1 denotes CEU ancestry and 2 denotes YRI. The estimates of these parameters for the three groups of CEU-YRI simulated samples are displayed in Table 3.7, along with the estimates of the unconditional misfitting probabilities found in the analysis of section 3.2.4.

We can see from these results that in general the conditional misfitting probabilities for which there is no misfitting at the preceding window ($m_{..11}$ and $m_{..22}$) are in line

r		$ij = 11$	$ij = 12$	$ij = 21$	$ij = 22$
5	independent of uv	0.964	0.0362	0.0386	0.962
5	$uv = 11$	0.965	0.0346	0.196	0.804
5	$uv = 12$	0.933	0.0672	0.769	0.231
5	$uv = 21$	1.000	0.000	0.130	0.870
5	$uv = 22$	0.574	0.426	0.0361	0.964
10	independent of uv	0.959	0.0414	0.0401	0.960
10	$uv = 11$	0.960	0.0400	0.0382	0.962
10	$uv = 12$	0.922	0.0781	0.000	1.000
10	$uv = 21$	1.000	0.000	0.117	0.883
10	$uv = 22$	0.750	0.250	0.0379	0.962
50	independent of uv	0.956	0.0436	0.0407	0.959
50	$uv = 11$	0.957	0.0430	0.0825	0.917
50	$uv = 12$	0.830	0.170	0.000	1.000
50	$uv = 21$	1.000	0.000	0.138	0.862
50	$uv = 22$	0.814	0.186	0.0404	0.960

Table 3.7: The EM-MULTIMIX estimates of the conditional misfitting probabilities, m_{ijuv}, for the CEU-YRI simulated samples for $r \in \{5, 10, 50\}$. The estimates of the unconditional misfitting probabilities, m_{ij}, which are independent of u and v are also shown.

with the corresponding estimates in the unconditional model. The exception to this was the estimates of m_{1122} and m_{1222} which, for each r-group of samples, were notably different from the m_{11} and m_{12} estimates respectively. More specifically, m_{1222} was larger than m_{12} (0.426 as compared to 0.0361 for $r = 5$) meaning that the probability of misfitting to population YRI when the underlying ancestry is CEU is greater than in the unconditional model if at the previous state the ancestry is YRI and there is no misfitting.

Where there is misfitting at the previous window, the probability of misfitting at the current window, when the underlying ancestral state is unchanged, is higher than the corresponding misfitting probability in the original model. That is, m_{1212} was estimated to be greater than m_{12}, and m_{2121} was estimated to be greater than m_{21}. This shows that misfitting extends across windows, suggesting that there are regions of the genome longer than a single window in length where the MVN model for the true

ancestral population does not fit the sample's data as well as for the alternative population.

3.3 Three-way admixture

An additional strength of our method is that it is directly applicable when the admixture involves more than two ancestral populations - an asset not possessed by HAPMIX. Being able to discern ancestry in a multi-way admixture scenario will be particularly useful when conducting disease association studies or admixture mapping in Latino populations, whose ancestry is a mixture of European, Native American and African (Salzano and Bortolini, 2002). This aspect of our model is therefore a useful feature in the analysis of real admixed populations.

To investigate how MULTIMIX performs in the case of three-way admixture we simulated 9 groups of CEU-YRI-CHB admixed individuals with 5 individuals in each group. The groups were specified by combinations of the r and (q_{YRI}, q_{CHB}) parameters, with $q_{CEU} = 0.4$ in all cases. The global proportions of YRI and CHB ancestry in the simulations were $(0, 0.6)$, $(0.1, 0.5)$ and $(0.3, 0.3)$ and we considered $r \in \{5, 10, 20\}$. These parameter values were chosen to emulate a range of recent admixture scenarios similar to those estimated for Hispanic and Latino populations including Puerto Ricans, Dominicans, Ecuadorians and Mexicans (Bryc et al., 2010b).

In analysing the CEU-YRI-CHB simulated individuals at chromosome 1 we considered values of $\lambda \in \{0.005, 0.05, 0.1\}$ and used a window size of 100 SNPs. The performance of each of the three implementations of MULTIMIX for the 9 parameter groups is reported in Tables 3.8-3.10. The misfitting probabilities were estimated in the EM and CEM implementations while in the MCMC runs they were kept fixed at the estimates obtained by the corresponding EM runs. As in the case of the CEU-YRI analysis, we did not find one method that outperformed the others across the nine parameter

groups. The total performance across all of the simulated samples, at the optimum value of λ for each group, was found to be 98.031% for the MCMC, 98.018% for the EM and 98.001% for the CEM method.

A comparison of the true ancestry of a chromosome simulated to be 40%, 10% and 50% CEU, YRI and CHB respectively, with $r = 10$ and the calls made by the three methods is displayed in Fig.3.4. Painting (a) displays the simulated ancestry that our model aims to infer, with each vertical bar being coloured to represent the population of origin of the allele at a single SNP. The ancestry calls made by the MCMC, EM and CEM methods are illustrated in (b), (c) and (d) respectively. It is clear that all methods perform well in this case, with the differences between the MCMC and EM estimates only occurring near the boundaries of ancestral chunks. The notable error of the CEM estimates is a very narrow segment of European ancestry that has been called as CHB. The MCMC and EM methods are more similar in that they both estimate the marginal posterior probabilities of each ancestry state in the windows and then the ancestry is classified according to which population is of the highest probability, while the CEM method makes ancestry calls by finding the jointly most likely set of calls which may not be the same as those of highest marginal probability.

In the EM and CEM runs, where the misfitting probabilities were estimated rather than fixed, we found that the probability of misfitting at a window where the ancestry is CEU is estimated to be higher than that at one where it is CHB, with both of these being higher than for YRI ancestry. That is, in general, $m_{11} < m_{33} < m_{22}$ where 1, 2 and 3 denote CEU, YRI and CHB respectively. This may be expected since the variation that is seen in European and Asian populations is to some extent a subset of that seen in Africa which is more genetically diverse, as explained by the Out-of-Africa hypothesis of modern human origins. This means that chunks of European ancestry are more likely to "look" like African segments, as are Asian chunks, than Africans would look like European or Asian chunks. The estimates of the misfitting probabilities seem to

r	$(q_{CEU}, q_{YRI}, q_{CHB})$	λ	Method	m_{11}	m_{12}	m_{13}	m_{21}	m_{22}	m_{23}	m_{31}	m_{32}	m_{33}	Performance (%)	resolved
5	(0.4 , 0.0 , 0.6)	0.005	MCMC										97.698	97.789
5	(0.4 , 0.0 , 0.6)	0.005	EM	0.788	0.0267	0.186	0.418	0.388	0.194	0.134	0.0336	0.832	98.485	98.531
5	(0.4 , 0.0 , 0.6)	0.005	CEM	0.789	0.0259	0.185	0.0250	0.950	0.0250	0.132	0.0333	0.835	98.487	98.519
5	(0.4 , 0.0 , 0.6)	0.05	MCMC										98.627	98.645
5	(0.4 , 0.0 , 0.6)	0.05	EM	0.843	0.00549	0.151	0.698	0.0251	0.277	0.0786	0.00758	0.914	96.081	96.161
5	(0.4 , 0.0 , 0.6)	0.05	CEM	0.841	0.00692	0.152	0.0250	0.950	0.0250	0.0752	0.00788	0.917	98.534	98.520
5	(0.4 , 0.0 , 0.6)	0.1	MCMC										98.661	**98.957**
5	(0.4 , 0.0 , 0.6)	0.1	EM	0.843	0.00549	0.151	0.698	0.0251	0.277	0.0786	0.00758	0.914	98.467	98.505
5	(0.4 , 0.0 , 0.6)	0.1	CEM	0.866	0.00348	0.130	0.0250	0.950	0.0250	0.0557	0.00402	0.940	98.409	98.432
5	(0.4 , 0.1 , 0.5)	0.005	MCMC										99.032	**99.169**
5	(0.4 , 0.1 , 0.5)	0.005	EM	0.779	0.0318	0.189	0.0323	0.947	0.0210	0.132	0.0369	0.831	99.026	99.164
5	(0.4 , 0.1 , 0.5)	0.005	CEM	0.780	0.0309	0.190	0.0315	0.948	0.0204	0.131	0.0381	0.831	98.972	99.090
5	(0.4 , 0.1 , 0.5)	0.05	MCMC										98.709	98.795
5	(0.4 , 0.1 , 0.5)	0.05	EM	0.826	0.00745	0.167	0.0290	0.928	0.0432	0.0696	0.0052	0.925	98.749	98.852
5	(0.4 , 0.1 , 0.5)	0.05	CEM	0.827	0.00748	0.166	0.0282	0.928	0.0435	0.0639	0.00544	0.931	98.683	98.772
5	(0.4 , 0.1 , 0.5)	0.1	MCMC										98.666	98.761
5	(0.4 , 0.1 , 0.5)	0.1	EM	0.836	0.00481	0.159	0.0304	0.924	0.0453	0.0552	0.00275	0.942	98.643	98.730
5	(0.4 , 0.1 , 0.5)	0.1	CEM	0.855	0.00515	0.140	0.0274	0.930	0.0428	0.0451	0.00235	0.953	98.591	98.693
5	(0.4 , 0.3 , 0.3)	0.005	MCMC										99.313	99.420
5	(0.4 , 0.3 , 0.3)	0.005	EM	0.803	0.0305	0.167	0.0312	0.942	0.0270	0.136	0.0365	0.828	99.298	**99.439**
5	(0.4 , 0.3 , 0.3)	0.005	CEM	0.806	0.0302	0.164	0.0307	0.943	0.0260	0.136	0.0367	0.827	99.301	99.419
5	(0.4 , 0.3 , 0.3)	0.05	MCMC										99.250	99.353
5	(0.4 , 0.3 , 0.3)	0.05	EM	0.845	0.00847	0.146	0.0367	0.918	0.0453	0.0751	0.00753	0.917	99.259	99.340
5	(0.4 , 0.3 , 0.3)	0.05	CEM	0.847	0.00912	0.144	0.0324	0.921	0.0469	0.0687	0.00692	0.924	99.298	99.380
5	(0.4 , 0.3 , 0.3)	0.1	MCMC										99.219	99.328
5	(0.4 , 0.3 , 0.3)	0.1	EM	0.861	0.00400	0.135	0.0404	0.908	0.0513	0.0635	0.00342	0.933	99.219	99.329
5	(0.4 , 0.3 , 0.3)	0.1	CEM	0.879	0.00447	0.117	0.0367	0.915	0.0482	0.0582	0.00381	0.938	99.074	99.173

Table 3.8: The performance of MULTIMIX on chromosome 1 of the simulated CEU-YRI-CHB admixed individuals for different values of λ. The model was fit by three different methods: MCMC, the EM algorithm and the CEM algorithm. The misfitting probabilities m were estimated in the EM and CEM runs, and in the MCMC method they were kept constant at the estimates obtained by the corresponding EM run.

r	$(q_{CEU}, q_{YRI}, q_{CHB})$	λ	Method	m_{11}	m_{12}	m_{13}	m_{21}	m_{22}	m_{23}	m_{31}	m_{32}	m_{33}	Performance (%)	resolved
10	(0.4 , 0.0 , 0.6)	0.005	MCMC										97.877	98.117
10	(0.4 , 0.0 , 0.6)	0.005	EM	0.790	0.0261	0.184	0.632	0.0840	0.284	0.141	0.0319	0.827	98.063	98.288
10	(0.4 , 0.0 , 0.6)	0.005	CEM	0.788	0.0279	0.184	0.0250	0.950	0.0250	0.140	0.0314	0.829	98.089	98.229
10	(0.4 , 0.0 , 0.6)	0.05	MCMC										98.198	98.260
10	(0.4 , 0.0 , 0.6)	0.05	EM	0.820	0.00532	0.175	0.818	0.0875	0.0948	0.0830	0.00583	0.911	97.312	97.440
10	(0.4 , 0.0 , 0.6)	0.05	CEM	0.826	0.00744	0.166	0.0250	0.950	0.0250	0.0794	0.00567	0.915	98.408	**98.488**
10	(0.4 , 0.0 , 0.6)	0.1	MCMC										97.770	97.967
10	(0.4 , 0.0 , 0.6)	0.1	EM	0.834	0.00486	0.161	0.498	0.119	0.383	0.0672	0.00193	0.931	97.896	98.007
10	(0.4 , 0.0 , 0.6)	0.1	CEM	0.857	0.00520	0.138	0.0250	0.950	0.0250	0.0608	0.00194	0.937	98.280	98.370
10	(0.4 , 0.1 , 0.5)	0.005	MCMC										97.936	98.154
10	(0.4 , 0.1 , 0.5)	0.005	EM	0.797	0.0286	0.174	0.0336	0.955	0.0119	0.124	0.0293	0.846	97.984	98.200
10	(0.4 , 0.1 , 0.5)	0.005	CEM	0.797	0.0289	0.174	0.0322	0.958	0.0103	0.122	0.0295	0.849	98.093	**98.335**
10	(0.4 , 0.1 , 0.5)	0.05	MCMC										98.100	98.250
10	(0.4 , 0.1 , 0.5)	0.05	EM	0.838	0.00937	0.153	0.0407	0.939	0.0199	0.0752	0.00387	0.921	98.091	98.245
10	(0.4 , 0.1 , 0.5)	0.05	CEM	0.847	0.00944	0.144	0.0374	0.942	0.0206	0.0695	0.00343	0.927	97.854	97.966
10	(0.4 , 0.1 , 0.5)	0.1	MCMC										97.899	98.000
10	(0.4 , 0.1 , 0.5)	0.1	EM	0.851	0.00578	0.144	0.0437	0.923	0.0329	0.0591	0.00186	0.939	97.940	98.058
10	(0.4 , 0.1 , 0.5)	0.1	CEM	0.870	0.00532	0.125	0.0388	0.934	0.0272	0.0474	0.00181	0.951	97.730	97.811
10	(0.4 , 0.3 , 0.3)	0.005	MCMC										97.861	98.005
10	(0.4 , 0.3 , 0.3)	0.005	EM	0.787	0.0261	0.187	0.0294	0.952	0.0189	0.137	0.0272	0.836	97.900	98.036
10	(0.4 , 0.3 , 0.3)	0.005	CEM	0.783	0.0260	0.191	0.0283	0.953	0.0185	0.135	0.0263	0.839	98.116	98.263
10	(0.4 , 0.3 , 0.3)	0.05	MCMC										98.203	**98.354**
10	(0.4 , 0.3 , 0.3)	0.05	EM	0.840	0.0067	0.153	0.0406	0.923	0.0366	0.0818	0.00357	0.915	98.199	98.343
10	(0.4 , 0.3 , 0.3)	0.05	CEM	0.846	0.00621	0.147	0.0392	0.925	0.0363	0.0800	0.00354	0.916	98.015	98.171
10	(0.4 , 0.3 , 0.3)	0.1	MCMC										98.115	98.266
10	(0.4 , 0.3 , 0.3)	0.1	EM	0.855	0.00469	0.141	0.0455	0.912	0.0427	0.063	0.00161	0.935	98.025	98.168
10	(0.4 , 0.3 , 0.3)	0.1	CEM	0.872	0.00433	0.124	0.0404	0.919	0.0409	0.0562	0.0014	0.942	97.798	97.973

Table 3.9: Table 3.8 continued.

r	$(q_{CEU}, q_{YRI}, q_{CHB})$	λ	Method	m_{11}	m_{12}	m_{13}	m_{21}	m_{22}	m_{23}	m_{31}	m_{32}	m_{33}	Performance (%)	resolved
20	(0.4 , 0.0 , 0.6)	0.005	MCMC										95.769	96.094
20	(0.4 , 0.0 , 0.6)	0.005	EM	0.802	0.0307	0.167	0.161	0.542	0.297	0.127	0.0317	0.841	96.077	**96.338**
20	(0.4 , 0.0 , 0.6)	0.005	CEM	0.812	0.0306	0.158	0.0250	0.950	0.0250	0.126	0.0321	0.842	96.017	96.280
20	(0.4 , 0.0 , 0.6)	0.05	MCMC										95.832	95.831
20	(0.4 , 0.0 , 0.6)	0.05	EM	0.857	0.00372	0.139	0.778	0.0500	0.172	0.0703	0.00434	0.925	93.351	93.537
20	(0.4 , 0.0 , 0.6)	0.05	CEM	0.868	0.00821	0.124	0.0250	0.950	0.0250	0.0643	0.00436	0.931	96.012	96.159
20	(0.4 , 0.0 , 0.6)	0.1	MCMC										95.635	95.868
20	(0.4 , 0.0 , 0.6)	0.1	EM	0.868	0.00162	0.131	0.800	0.0466	0.153	0.0547	0.00163	0.944	93.706	93.890
20	(0.4 , 0.0 , 0.6)	0.1	CEM	0.887	0.00607	0.107	0.0250	0.950	0.0250	0.0460	0.00170	0.952	95.942	96.084
20	(0.4 , 0.1 , 0.5)	0.005	MCMC										95.897	96.256
20	(0.4 , 0.1 , 0.5)	0.005	EM	0.814	0.0242	0.162	0.0310	0.946	0.0227	0.131	0.0363	0.832	95.914	96.310
20	(0.4 , 0.1 , 0.5)	0.005	CEM	0.819	0.0252	0.155	0.0265	0.955	0.0180	0.138	0.0365	0.826	95.536	95.808
20	(0.4 , 0.1 , 0.5)	0.05	MCMC										96.659	**96.810**
20	(0.4 , 0.1 , 0.5)	0.05	EM	0.860	0.00720	0.133	0.0367	0.920	0.0429	0.0741	0.00696	0.919	96.612	96.744
20	(0.4 , 0.1 , 0.5)	0.05	CEM	0.871	0.00685	0.122	0.0343	0.929	0.0364	0.0677	0.00822	0.924	96.446	96.634
20	(0.4 , 0.1 , 0.5)	0.1	MCMC										96.590	96.702
20	(0.4 , 0.1 , 0.5)	0.1	EM	0.875	0.00379	0.122	0.0441	0.906	0.0501	0.0572	0.00351	0.939	96.632	96.752
20	(0.4 , 0.1 , 0.5)	0.1	CEM	0.895	0.00371	0.102	0.0356	0.919	0.0453	0.0489	0.00428	0.947	96.465	96.588
20	(0.4 , 0.3 , 0.3)	0.005	MCMC										95.932	96.441
20	(0.4 , 0.3 , 0.3)	0.005	EM	0.795	0.0288	0.176	0.027	0.948	0.0246	0.146	0.0348	0.819	95.968	96.435
20	(0.4 , 0.3 , 0.3)	0.005	CEM	0.795	0.0286	0.176	0.0258	0.951	0.0228	0.143	0.0350	0.822	96.064	96.534
20	(0.4 , 0.3 , 0.3)	0.05	MCMC										96.620	96.965
20	(0.4 , 0.3 , 0.3)	0.05	EM	0.842	0.00733	0.151	0.033	0.929	0.0376	0.0898	0.00732	0.903	96.686	**97.065**
20	(0.4 , 0.3 , 0.3)	0.05	CEM	0.851	0.00686	0.142	0.0316	0.934	0.0339	0.0833	0.00754	0.909	96.560	96.892
20	(0.4 , 0.3 , 0.3)	0.1	MCMC										96.672	96.950
20	(0.4 , 0.3 , 0.3)	0.1	EM	0.853	0.0052	0.142	0.0342	0.921	0.0447	0.0655	0.00437	0.930	96.715	97.006
20	(0.4 , 0.3 , 0.3)	0.1	CEM	0.870	0.00565	0.124	0.0276	0.931	0.0412	0.0537	0.00488	0.941	96.476	96.764

Table 3.10: Table 3.8 continued.

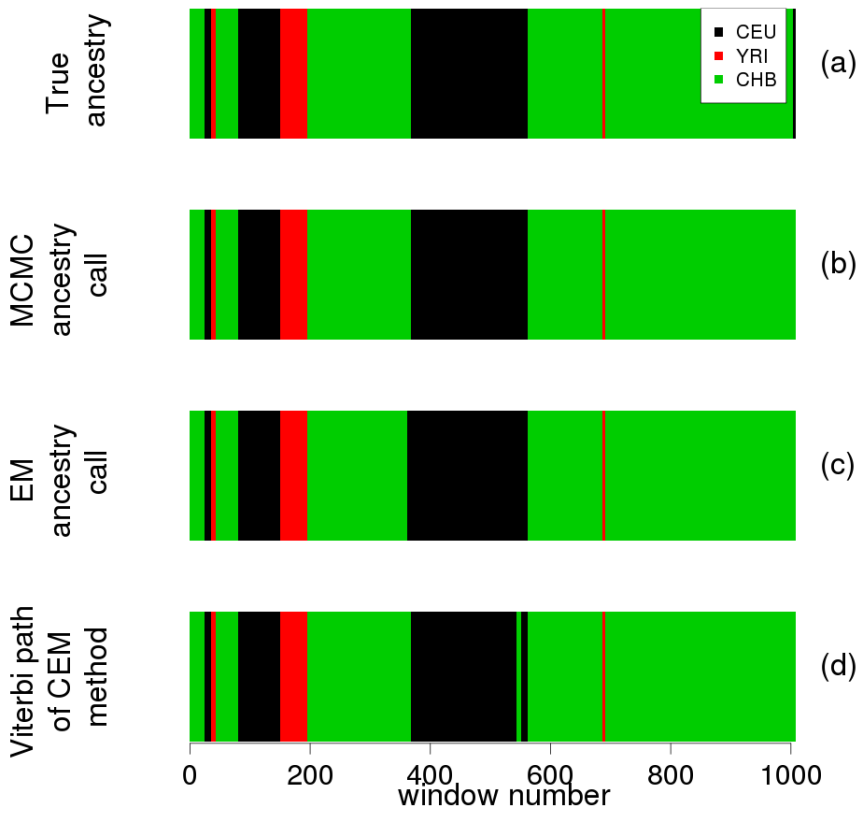

Figure 3.4: (a) The true ancestry of a single copy of chromosome 1 simulated with parameters r=10 and $(q_{CEU}, q_{YRI}, q_{CHB}) = (0.4, 0.1, 0.5)$. (b) MCMC-MULTIMIX ancestry call. (c) Ancestry call by EM-MULTIMIX. (d) Viterbi path of the CEM-MULTIMIX method

reflect this relationship between the three populations.

3.4 Five-way admixture

With the aim of testing how well our model can discern ancestry chunks when the admixed individual has ancestry from many different populations, we tested it in a five-way setting. While this may not be representative of most real admixed groups, it gives us an idea of the extent to which our model can handle a very complicated admixture scenario. We simulated one individual to be five-way admixed from the CEU, YRI, CHB, GIH and collective Native American populations at 582,186 SNPs genome wide that are common to both the HapMap and HGDP data sets. We set $r = 10$

Figure 3.5: A comparison of the true and inferred ancestry along one copy of chromosome 2 for an individual simulated to be of equal proportions of CEU, YRI, CHB, GIH and Native American ancestry for $r = 10$.

and set the global proportions of each ancestry to be equal at 20% each. The ancestry of this individual was inferred via the EM method of MULTIMIX with a window size of 100 SNPs and $\lambda = 0.005$, where the miscopying probabilities were estimated.

When applied to the five-way simulated individual, the EM method of MULTIMIX attained an accuracy of 91.456% of local ancestry calls. A comparison of the simulated and inferred ancestry of one copy of chromosome 2 is illustrated in Fig.3.5 where we can see that despite the elaborate admixture pattern of this sample, our method is still a very powerful tool in ascertaining the correct ancestral population along the chromosome. Broadly MULTIMIX succeeds in identifying the different chunks of ancestry, and infers the switch points to be very near to where they actually occur.

A summary of the percent of calls made to be of each population, for each true ancestry is reported in Table 3.11. It is the YRI sites that are most correctly identified which may be explained by the fact that this population is more diverged from the others. Interestingly, it is the CEU chunks that were most poorly ascertained, with over 12% of CEU sites being called as GIH. One possible explanation for this is that the GIH samples may in fact contain individuals who have some extent of European ancestry although this should be further investigated.

	CEU	YRI	CHB	GIH	Native American
CEU	83.679	0.482	2.000	12.533	1.306
YRI	0.501	97.916	0.466	0.685	0.433
CHB	1.661	0.535	93.311	3.160	1.333
GIH	5.040	0.405	1.602	90.660	2.294
Native American	1.750	0.712	2.278	2.390	92.871

Table 3.11: Percent of calls made in the analysis of a simulated 5-way admixed individual, where the true ancestry is displayed in the row and the ancestry call is in the column.

3.5 Summary

In this chapter we assessed the performance of our MULTIMIX model via simulations in which admixed chromosomes were generated for two-way and three-way mixing for different ages of admixture. The MCMC implementation was first studied to find the optimal values of the user-specified parameters in the case of European and African admixture. This analysis suggested using a window size of 100 SNPs, setting λ equal to 0.005 and using a misfitting probability of 0.05 gave the highest accuracy of calls at 99.225%. Applying the EM and CEM methods to this same set of two-way samples showed that none of the three methods comprehensively outperformed the other two and that they each offer an accurate tool to make local ancestry calls. The EM and CEM methods were found to be faster to implement than the MCMC sampling approach, and furthermore were more than twice as fast as HAPMIX. The accuracy of the MULTIMIX calls in the CEU-YRI samples was slightly lower than that achieved by HAPMIX, being only 0.118% and 0. 174% less for the samples generated for 5 and 10 generations of mixing, respectively, which is a realistic range of age of admixture expected to be seen in African American individuals. This shows that MULTIMIX is competitive with the leading methods in the field.

To test how sensitive our model was to violations of the assumption that the hap-

lotypes in the panels are representative of the ancestors of the admixed individual, we ran MCMC-MULTIMIX on the CEU-YRI samples substituting the appropriate ancestral panels for those of a closely-related but different population, namely the Toscani Italians (TSI) and the Maasai Kenyans (MKK). Using the TSI panel in place of the CEU panel had the least effect on the performance of both MULTIMIX and HAPMIX, while replacing the MKK for the YRI lead to the least accurate ancestry calls of both methods however in this set-up it was MULTIMIX that outperformed HAPMIX.

A strength of our MULTIMIX model is that it is applicable in a variety of scenarios where both the admixed samples and the ancestral panels may be phased or unphased. MULTIMIX is unique in that it can take as input an unphased panel of ancestral genotypes at a dense-set of linked SNPs, a feature that is particularly useful if an accurately phased panel of ancestral haplotypes is not available. Testing MCMC-MULTIMIX on the CEU-YRI admixed samples in the four possible phasing settings, we found that while phasing of both the panel and samples gave the most correct calls, using an unphased panel with phased admixed samples reduced performance by only 0.013%. There was a larger drop in performance of just over 1% when the sample was not phased, but these results show that the model is useful in all four phasing settings that may arise in the analysis of real data.

Since real admixed populations such as Latino and Hispanic groups may exhibit ancestry inherited from more than two continents, we simulated three-way admixed samples of CEU, YRI and Han Chinese (CHB) ancestry for a range of global proportions and ages of admixture selected to be representative of real admixed individuals. The MCMC, EM and CEM methods were compared across the nine different parameter groups to reveal that in different groups different methods gave the most accurate calls and that all three of the methods achieved an over all performance of just over 98% in these simulations.

Finally, to test the limits of how well our model is able to accurately ascertain lo-

cal ancestry in a multi-way admixture setting, we applied EM-MULTIMIX to a single individual simulated to possess ancestry from five different populations. We found that the over all performance was just over 91% and therefore considerably lower than in the three-way setting tested previously, but the general pattern of ancestral chunks was uncovered by the model. There was a notable difference in accuracy depending on the true ancestral population. That is, nearly 98% of YRI sites were correctly identified but as low as 83.7% of CEU sites were called without error, most mistakes occurring as GIH calls. While admixture involving this number of populations may be unrealistic or rare, it demonstrates that MULTIMIX can handle the computational load of five candidate ancestral populations and that its performance is above that which may be expected in such a challenging setting.

In the next chapter we compare MULTIMIX to two other methods of local ancestry estimation - LAMP-LD and RFmix - as well as HAPMIX, in a simulation study to judge the performance of these four techniques on both two-way and three-way admixed individuals.

Chapter 4

Comparison of MULTIMIX with other methods

In Chapter 3 we performed a simulation study to compare the performance of MULTIMIX with HAPMIX in the context of two-way admixture. In this chapter, we make a more extensive comparison of our method to RFmix and LAMP-LD, as well as HAPMIX, in both two and three-way admixture simulations. In these simulations, both the ancestral panels and the admixed samples are phased. We look at how accurately each of the four methods can call local ancestry and how similar the estimates of the different methods are. Furthermore, we examine how frequently the various types of errors are made by each method and we assess how accurately they are able to ascertain the precise location of a switch in ancestry.

4.1 Simulating the Admixed Samples

In this study, both two and three-way admixed samples were simulated and all simulated data was produced by members of the Bustamante[1] group. This work was con-

[1]Carlos D. Bustamante, Department of Genetics, Stanford University, School of Medicine.

ducted to be a comparison of ancestry estimation methods as part of the 1000 Genomes Project. The two-way admixture involved the European CEU and the African YRI populations of HapMap, for which 20 haploid admixed individuals were simulated across all 22 autosomes. These samples were simulated to have global ancestry proportions that were 82% YRI and 18% CEU, for 8 generations of mixing. These particular values of ancestry proportions and age of admixture were chosen to be representative of admixture in African Americans and this study therefore gives us an indication of how well each of the four methods would perform on real data.

One admixed haplotype was simulated at a time, first by assigning the ancestry at each locus for the specified number of generations of mixing and global ancestry proportions. The haplotypes are then filled in by selecting one haplotype at random from a set of 10 phased trios belonging to the appropriate population. The selected haplotype is used to fill in the haplotype information of the admixed sample along the length of the chromosome at sites with the appropriate ancestry. The individuals that were used in simulating the admixed samples were then set aside and did not form part of the ancestral panels.

For the two-way simulations, the ancestral panels were composed of 184 CEU and 184 YRI haplotypes, typed at 663,652 autosomal biallelic Omni SNPs. We stress that the haplotypes in these ancestral panels are separate to those from which the admixture samples were generated. The samples were also typed at these same SNPs. In the following analysis, we will refer to these two-way admixed samples as the Omni ASW2.

The three-way admixture was between the European CEU, the African YRI and a Native American (NAH) population. As in the two-way scenario, 20 haploid admixed samples were simulated across all autosomes. In this scenario, the samples were simulated to be predominantly of CEU and NAH ancestry, with a small YRI component with the global ancestry proportions set to 5% YRI, 50% CEU and 45% NAH. The sim-

ulations were conducted for 12 generations of mixing. As in the two-way simulations described above, the three-way simulations were selected to be representative of real admixture, in this case that of some Latino populations. Each of the three ancestral panels contained 66 phased haplotypes. The admixed samples were also phased and typed at the same 614,209 Affy SNPs as the panels. The three-way simulations will be referred to as the Affy MXL3 samples.

The following people were involved in carrying out the analysis of the admixed samples by the three other competing methods: **RFmix** - Eimear E. Kenny, Brian Maples; **LAMP-LD** - Yael Baran; **HAPMIX** - Anjali Gupta Hinch, Amy Williams.

4.2 Omni Two-way Admixture Simulations

In these simulations, as with the three-way admixture, we analysed the samples using our MULTIMIX method, the other three methods being applied by their corresponding authors. The MULTIMIX results reported here were the best of those found by the MCMC, EM and CEM methods, with the step to resolve the boundaries between ancestral chunks. For the MULTIMIX and HAPMIX analyses, the marginal posterior probability of all possible ancestry states at each SNP was available, and the state with the highest probability was taken to be the ancestry call at that site. The RFmix and LAMP-LD results consisted simply of the ancestry calls at each SNP rather than their posterior probabilities. Performance was measured as the percentage of sites at which the ancestry call was correct, a call being made at every SNP in the data set.

4.2.1 Performance over each ancestry

Given the ancestry calls of the four methods for each of the admixed samples, we found that they all achieved a very high performance, as would be expected in this scenario due to the CEU and YRI populations being well diverged. HAPMIX made the most

	% of CEU sites called as CEU	% of CEU sites called as YRI	% of YRI sites called as YRI	% of YRI sites called as CEU	% of correct calls
RFmix	97.044	2.956	99.302	0.698	98.955
LAMP-LD	99.278	0.722	99.729	0.271	99.659
MULTIMIX	99.155	0.845	99.749	0.251	99.658
HAPMIX	99.351	0.649	99.827	0.173	99.754

Table 4.1: The percentage of correct and incorrect ancestry calls for each simulated ancestry, CEU and YRI, of the four methods applied to the Omni ASW2 data set.

correct ancestry calls at 99.754%, closely followed by LAMP-LD and MULTIMIX with 99.659% and 99.658% correct respectively. The RFmix method gave the least number of correct calls at 98.955%.

Stratifying performance by the true ancestral population, we saw that for each method a slightly higher proportion of YRI sites were called correctly than were the CEU sites as shown in Table 4.1. This difference, illustrated in Fig. 4.1, which was approximately 0.5%, may be explained by the fact that the majority of ancestry in the samples of YRI meaning that shorter stretches of ancestry, which are harder to call correctly than longer chunks, will tend to be of CEU origin. The tendency of each method to commit errors at narrow chunks is investigated fully in section 4.2.5.

4.2.2 Agreement between methods

In order to quantify how similar the local ancestry calls of the four methods were, for each pair of methods we looked at the number of sites at which the calls agreed or differed. The proportion of calls that are the same between each pair of methods is reported in Table 4.2. In general, there is very good agreement between all of the methods, above 98.9% in each instance, reflecting the fact that European-African admixture is relatively simple to ascertain and all methods gave highly accurate results.

The calls made by HAPMIX and those made by MULTIMIX were the most similar

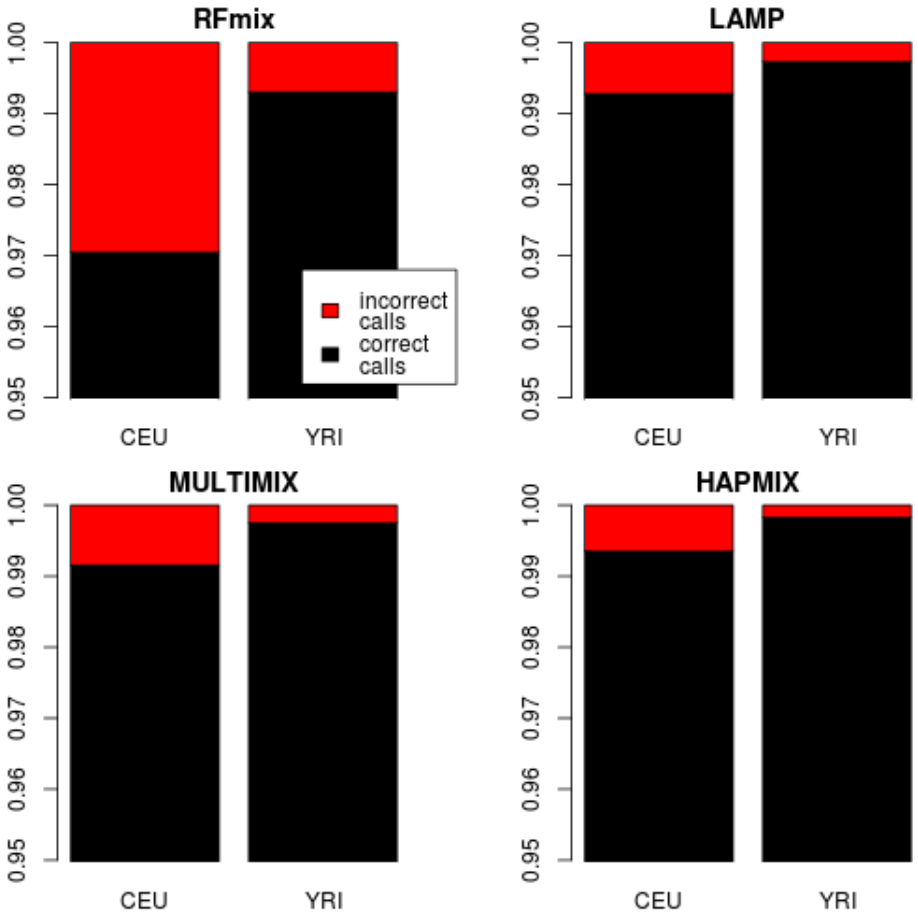

Figure 4.1: This figure displays the proportion of incorrect calls of each ancestry for the four methods in the Omni ASW2 simulations.

	LAMP-LD	MULTIMIX	HAPMIX
RFmix	98.933	98.999	98.981
LAMP-LD		99.636	99.703
MULTIMIX			99.712

Table 4.2: The percentage of calls that agree between each pair of methods in the Omni ASW2 analysis.

across the methods tested, with 99.712% of their calls being in agreement. Calls made by LAMP-LD were also in very good agreement with those made by HAPMIX and MULTIMIX, being the same at 99.703% and 99.636% of sites respectively. The local ancestry inference of the RFmix method stood out as differing the most from the three other methods, with only 98.981% of calls being the same as those made by HAPMIX for example.

4.2.3 Comparison of errors between methods

To learn more about how the local ancestry calls of the methods compared, we examined how similar the errors made by each method were. To do this, we identified the sites at which a particular method made an incorrect ancestry call and asked, of these sites, how many were also assigned an incorrect ancestry call by each of the other techniques. This gave us an idea of the relative short-comings of each model.

The percentage of errors made by one method that are also errors in the other methods is displayed in Table 4.3. A high percentage reported between method a and method b, say, means that many of the errors in the former are also mistakes in the latter, indicating that the sites that are problematic for a are also wrongly inferred by b. Furthermore, if the percentage of errors made by method b that are also errors of method a is considerably lower in comparison, then this indicates a superiority of a over b.

The HAPMIX method had the highest percentage of shared errors over all, the most

	RFmix	LAMP-LD	MULTIMIX	HAPMIX
RFmix		15.248	18.493	13.044
LAMP-LD	46.855		46.819	42.538
MULTIMIX	56.473	46.527		43.881
HAPMIX	55.437	58.831	61.070	

Table 4.3: The percentage of errors made by the method listed in the row that are also errors made by the method listed in the column.

being with MULTIMIX at just over 61%. A smaller fraction of around 44% of the incorrect calls made by MULTIMIX were also made by HAPMIX, highlighting that there are more sites at which MULTIMIX is incorrect and HAPMIX is correct, than vice versa.

Once again, the inference of RFmix stands out as differing from that of the other three methods. There was a much smaller overlap in the fraction of wrongly called sites between RFmix and any other technique, with only approximately 15%, 18% and 13% of errors also being mistakes made by LAMP-LD, MULTIMIX and HAPMIX respectively. In addition, a high percentage of the errors made by the other methods were also made by RFmix, at around 46 - 55%, suggesting that RFmix not only goes wrong where other methods do not but that it also falls at the same hurdles as the other methods.

Having looked at how the local ancestry inferred by each method compared with that of the simulations, we have seen that over 98% of sites are correctly called by all four techniques (Table 4.1). Figure 4.2 shows a typical example of how the ancestry estimates of each method compare with each other and with the true simulated ancestry across a single chromosome copy. To shed more light on the differences in performance between one method and another, we investigated their susceptibility to the three types of errors that were found to occur upon examining chromosome paintings such as that of Fig.4.2.

The first occurs at the boundaries between consecutive chunks of differing ancestral populations. While all four ancestry callers may determine that there is a change

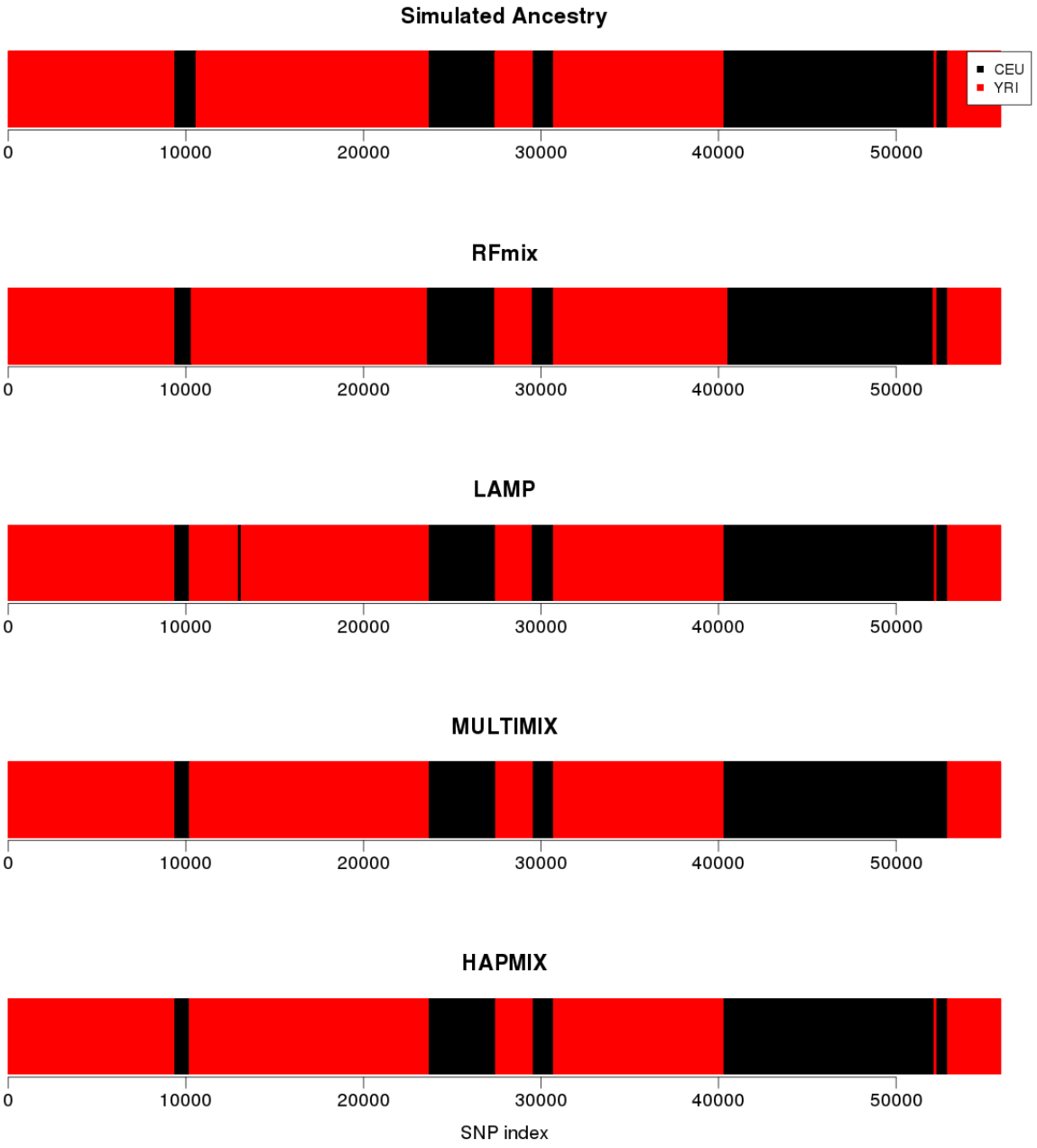

Figure 4.2: The ancestry calls by each of the four methods are displayed against the true simulated ancestry of a ASW2 sample at chromosome 1.

in ancestry within some range of the boundary, the ability of a method to accurately uncover the precise location of the switch will be advantageous, particularly in applications such as constructing recombination maps in admixed populations (section 1.6).

The second type of error concerns the ability of a method to correctly infer very short stretches of ancestry. We have seen in these simulations that there exist ancestral chunks of only a few hundred SNPs in length, and these may pose difficulties for the ancestry callers if they are not sensitive enough to detect them. If a method is missing these narrow chunks then it will not give an accurate picture of the amount of switching taking place and will therefore portray an admixture that is more recent than is really the case.

The third type of error also relates to short ancestral chunks. In this case however it is the inference of spurious switches in ancestry that is problematic. Here, a method that tends to call narrow changes in ancestry where they do not exist will infer a higher rate of switching and, conversely to the second type of error, will suggest the admixture to be older than it is. In the following sections we thoroughly examine how RFmix, LAMP-LD, MULTIMIX and HAPMIX compare across these three types of errors.

4.2.4 Accuracy of inferring boundaries

We sought to quantify how well each method was able to detect the precise location of a boundary between two chunks of different ancestry. We considered boundaries between chunks that were at least 1000 SNPs in length, of which there were 745 such chunks within the 20 admixed samples. At each of these sites, we counted the number of SNPs between the true location of the boundary in the simulated ancestry and the position at which it was inferred by each method, telling us how far off the called switch was from the true position. We did this for every sample across chromosomes

1 to 22.

We found that the HAPMIX method was able to infer the location of the boundaries most accurately, being on average about 10 SNPs off the precise location of a switch. MULTIMIX came in second, with its inferred boundaries being on average 14.8 SNPs away from the true site of the change. The performance of LAMP-LD was very similar, being off by an average of 15.2 SNPs. RFmix fared badly in this comparison as it made incorrect calls at approximately 77 SNPs per boundary site, showing it to be the poorest method of the four in ascertaining exactly where the ancestry switches occur.

4.2.5 Inferring short ancestral chunks

Having identified that the extent to which a method can correctly detect very short stretches of ancestry is a distinguishing factor between the different approaches, we sought to determine which method does this most accurately. From each of the simulated samples, we counted the number of SNPs that composed each ancestry chunk and focused only on those that were shorter than 1000 SNPs. We then examined these sites in the inferred ancestries of each of the four methods, and asked at how many of the SNPs in each chunk was the ancestry call correct. If the ancestry call disagreed with the true ancestry at all SNPs within the short chunk then the method had missed this stretch completely.

Of the 20 simulated haploid samples, there were 1727 ancestral chunks within the 22 autosomes. Of these, 366 chunks were of at most 1000 SNPs in length, totaling 165,520 SNPs. We focused exclusively on these SNPs to see which method showed the best sensitivity to detecting the presence of very narrow regions of changes in ancestry. We found that LAMP-LD was able to most accurately call the ancestral population at these sites, with 96.402% of these SNPs having correct calls. HAPMIX showed a similar performance at 95.494% and MULTIMIX attained 90.105%. Again, the RFmix method

stands out as being considerably poorer than the other three, with only 84.307% of SNPs in short chunks being called with the right ancestry.

While the ability to correctly ascertain very small stretches of ancestry is a strength in local ancestry estimation, the method will be weakened if it tends to infer short chunks spuriously. While this shortcoming may not lead to a large proportion of sites being called incorrectly, it will suggest more switching than in fact exists along a chromosome and consequently point to the admixture event being older than it really is. To compare this property across the four methods, we counted in how many instances an ancestral chunk of size less than 1000 SNPs was inferred but did not in fact exist in the true simulated ancestry. More specifically, if the ancestry called differed from the truth at all SNPs within the inferred narrow chunk then this was considered to be an instance of the third error type described earlier.

Across the 20 simulated samples, we found that MULTIMIX made the fewest of these errors, with only 32 such falsely inferred switches occurring. Nearly twice as many instances were seen in the HAPMIX analysis (61 spurious chunks) and there were 116 and 123 instances of this type of error by the LAMP-LD and RFmix methods respectively. The total number of SNPs incorrectly called in this manner by RFmix, LAMP-LD, MULTIMIX and HAPMIX were 35238, 23348, 13588, and 15256 respectively.

A summary of the number of SNPs at which the different kinds of errors were committed is reported for each of the four methods in Table 4.4. From this we see that for our MULTIMIX method, the error that lead to the highest number of SNPs being incorrectly called was where the method missed very narrow ancestry chunks. For both LAMP-LD and HAPMIX, it was instances in which spurious short stretches of ancestry were inferred that lead to the most incorrect calls. RFmix, whose performance was notably inferior to that of all other methods in each of the contexts examined in this analysis, made the highest number of incorrect calls at boundaries as compared to the number of errors at narrow chunks.

	RFmix	LAMP-LD	MULTIMIX	HAPMIX
Boundary errors	57,139 (0.430%)	11,351 (0.0855%)	11,057 (0.0833%)	7,081 (0.0533%)
Errors at short chunks	25,975 (0.196%)	5,956 (0.0449%)	16,378 (0.123%)	7,459 (0.0562%)
Spurious short chunks	35,238 (0.265%)	23,348 (0.176%)	13,588 (0.102%)	15,256 (0.115%)
Total errors	138,703 (1.045%)	45,261 (0.341%)	45,393 (0.342%)	32,651 (0.246%)

Table 4.4: Summary of the number of SNPs (and percentage of the total number of SNPs) at which the different types of errors were committed by each method in the Omni ASW2 simulations.

4.3 Affy Three-way Admixture Simulations

As for the Omni ASW2 simulations, we investigated the performance of RFmix, LAMP-LD, MULTIMIX and HAPMIX on the three-way Affy MXL3 samples, and assessed how often the different types of errors were committed for each of these. These simulations presented a more challenging scenario for the methods than the two-way case not only because the admixture involves more populations but because it includes a Native American group, who are less genetically diverged from Europeans than are Africans.

4.3.1 Performance over each ancestry

As we would expect, the over all performance of each method, which was measured as the percentage of sites at which the ancestry calls were correct, was lower in the MXL3 simulations than that seen in the ASW2 simulations. The relative performance of each method was the same as in the previous simulations, with HAPMIX making 97.561% of calls correctly, closely followed by LAMP-LD 96.919% and MULTIMIX 96.787%. The RFmix method did considerably worse getting only 94.626% of sites correct.

When we looked at the performance for each of the ancestral populations separately

	CEU called CEU	CEU called YRI	CEU called NAH	YRI called CEU	YRI called YRI	YRI called NAH	NAH called CEU	NAH called YRI	NAH called NAH	Over-all performance
RFmix	95.561	0.482	3.957	2.769	95.801	1.430	5.952	0.403	93.645	94.626
LAMP-LD	97.865	0.484	1.651	0.970	98.681	0.349	3.816	0.295	95.890	96.919
MULTIMIX	96.877	0.185	2.938	2.006	97.002	0.992	3.121	0.193	96.687	96.787
HAPMIX	97.607	0.310	2.083	0.854	98.686	0.460	2.242	0.313	97.444	97.561

Table 4.5: The percentage of sites of each true ancestry (CEU, YRI or NAH) that are called to be of each population, for the four methods applied to the Affy MXL3 data set.

which is displayed in Fig.4.3. We found the YRI sites were the most accurately inferred and the NAH sites were the hardest to discern (Table 4.5). This was true for all of the methods tested here. The most common incorrect ancestry call was where NAH sites were estimated to be CEU, which was the case in nearly 6% of calls made by RFmix and just over 2% of those made by HAPMIX. As mentioned previously, the NAH are less genetically diverged from the CEU than the YRI are, making it harder for any method to distinguish their haplotypes.

Furthermore, the Native American individuals themselves may already be admixed, possessing some degree of European or African ancestry. If this is the case, then regions of, say, European ancestry that exist in the Native American samples may be correctly identified by a method as being European but the simulated ancestry would be classified as Native American and so we would consider it to be an incorrect call. We did not investigate the possibility of admixture within the Native American samples themselves, but this may be done by performing a PCA analysis along with the CEU and YRI samples to see whether or not the Native American individuals form a well-defined cluster.

4.3.2 Agreement between methods

In the ASW2 simulations we found that the highest agreement of calls for any pair of methods was between HAPMIX and MULTIMIX. Here, however, we found that

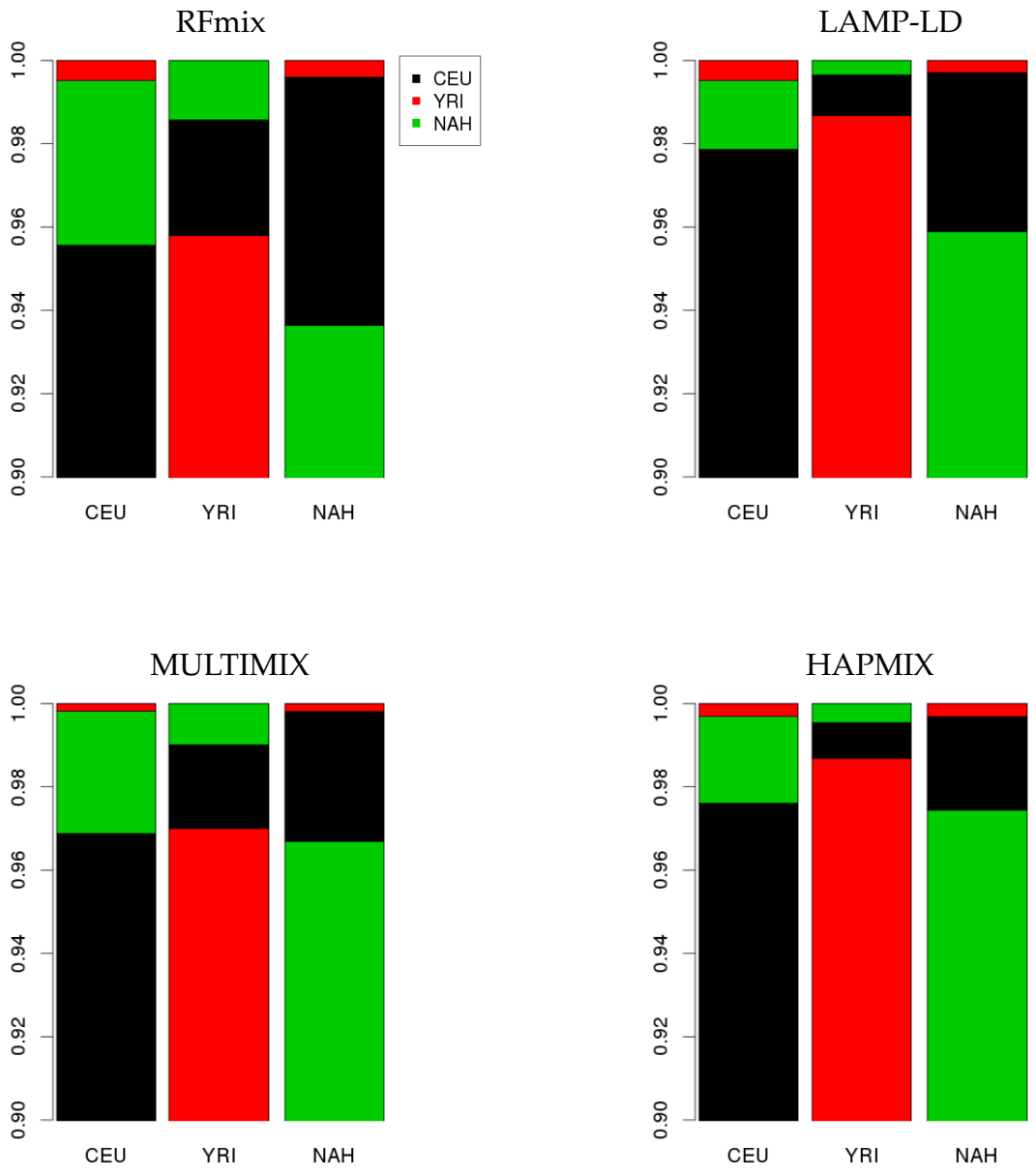

Figure 4.3: Proportion of calls of each population (CEU, YRI and NAH) stratified by the true simulated ancestry for the four methods.

	LAMP-LD	MULTIMIX	HAPMIX
RFmix	94.756	95.215	95.103
LAMP-LD		97.007	97.583
MULTIMIX			97.438

Table 4.6: The percentage of calls that agree between each pair of methods in the Affy MXL3 analysis.

	RFmix	LAMP-LD	MULTIMIX	HAPMIX
RFmix		30.505	35.780	27.530
LAMP-LD	53.203		54.240	51.117
MULTIMIX	59.849	52.022		48.677
HAPMIX	60.650	64.571	64.112	

Table 4.7: The percentage of errors made by the method listed in the row that are also errors made by the method listed in the column in the analysis of the Affy MXL3 simulations.

it was between HAPMIX and LAMP-LD at 97.583% agreement, although MULTIMIX did show similar agreement with HAPMIX at 97.438% (Table 4.6). Again, the RFmix method stood out as differing from the other three tested, showing approximately only 95% agreement with the other calls.

4.3.3 Comparison of errors between methods

We calculated the percentage of errors made by each method that are also errors in each of the other methods and these values are displayed in Table 4.7. HAPMIX had the highest proportion of shared errors with the other three, as was seen previously in the two-way ASW2 simulations. Also consistent with the ASW2 results was the observation that RFmix shared the smallest fraction of its errors with the other methods, indicating that it makes bad calls where other methods are able to infer the ancestral state correctly.

4.3.4 Accuracy of inferring boundaries

We have seen in the ASW2 simulations that the ability to accurately discern the location of a boundary between two stretches of differing ancestry is a distinguishing factor in the relative performance of the various methods. The number of SNPs at which the ancestry was incorrectly called due to boundary errors was higher for all four methods in the three-way simulations than in the two-way setting, which is to be expected as there is more switching in the MXL3 admixed samples so there are more boundaries.

As for the ASW2 simulations, we found that HAPMIX was best at identifying where the switches occurred with 0.795% of sites being incorrectly called around boundaries, followed by LAMP-LD and then MULTIMIX whose boundary errors happened at 0.862% and 1.141% of SNPs respectively (Table 4.8). RFmix made the most errors around the sites of boundaries with 2.3% of the total number of SNPs being called incorrectly in this way.

4.3.5 Inferring short ancestral chunks

In Fig.4.4 we show how the local ancestry estimates of RFmix, LAMP-LD, MULTIMIX and HAPMIX compare to each other and to the true simulated ancestry at a single copy of chromosome 1 in one of the MXL3 samples. We see that there are very short stretches of ancestry in the sample, of less than 1000 SNPs in length, and there are differences between the methods in being able to detect these. Across the 20 haploid samples there were 1,569 such narrow chunks encompassing a total of 669,337 SNPs. HAPMIX called approximately 90% of these sites correctly, closely matched by LAMP-LD with 89% and MULTIMIX with 85% correct. Again, RFmix stands out as being notably less accurate than the other methods as it only called 76% of SNPs in these narrow chunks correctly.

It is also apparent from Fig.4.4 that there are instances where some of the methods

	RFmix	LAMP-LD	MULTIMIX	HAPMIX
Boundary errors	282,503 (2.300%)	105,901 (0.862%)	140,146 (1.141%)	97,685 (0.795%)
Errors at short chunks	159,721 (1.300%)	73,132 (0.595%)	101,959 (0.830%)	66,766 (0.544%)
Spurious short chunks	183,304 (1.492%)	192,468 (1.567%)	123,532 (1.006%)	114,993 (0.936%)
Total errors	660,135 (5.374%)	378,509 (3.081%)	394,653 (3.213%)	299,645 (2.439%)

Table 4.8: Summary of the number of SNPs incorrectly called (and what percentage this is of the total number of SNPs called) due to the different types of errors committed by each method in the Affy MXL3 simulations.

mistakenly infer short changes in ancestry that do not in fact exist. MULTIMIX and HAPMIX were the least susceptible to errors in this regard, with only around 1% of sites being incorrectly called due to spurious switches (Table 4.8). The inference of LAMP-LD suffered the most from this error, its local ancestry estimates suggesting 703 incorrect short chunks to the 409 by RFmix, the 355 by HAPMIX and only 252 made by MULTIMIX.

4.4 Combining calls across methods

With four methods of local ancestry estimation available, we sought to determine if combining the calls from different methods led to an improvement in performance over that of each method separately. First we called the ancestry at each locus based upon a majority vote between the different combinations of three of the four methods. In the ASW2 simulations, since there are only two possible ancestries at any locus, it was possible to call 100% of sites by majority vote between three methods at a time. From Table 4.9 we see that the most correct calls were obtained when combining estimates from the LAMP-LD, MULTIMIX and HAPMIX methods, which increased

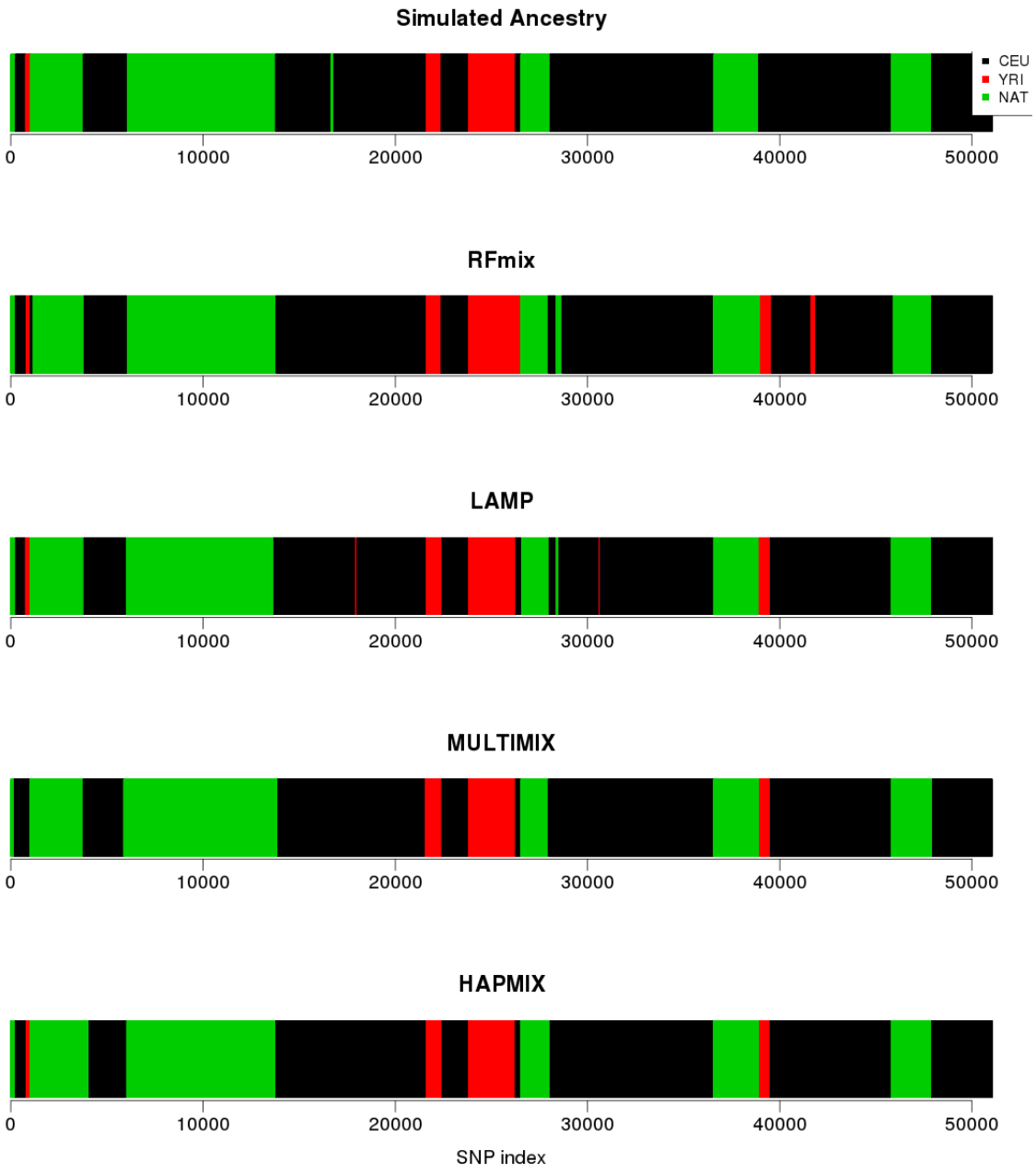

Figure 4.4: The ancestry calls by each of the four methods are displayed against the true simulated ancestry of a MXL3 sample at chromosome 1.

Methods used	ASW2 performance	% sites called	MXL3 performance	% sites called
LAMP-LD, MULTIMIX, HAPMIX	99.770	100	97.683	99.967
RFmix, MULTIMIX, HAPMIX	99.742	100	97.424	99.951
RFmix, LAMP-LD, HAPMIX	99.755	100	97.610	99.943
RFmix, LAMP-LD, MULTIMIX	99.700	100	97.216	99.952
RFmix, LAMP-LD, MULTIMIX, HAPMIX	99.915	98.712	98.964	92.989

Table 4.9: Performance achieved in the ASW2 and MXL3 simulations by combining calls across the different methods and the percent of sites at which a combined call could be made.

performance from the 99.754% obtained by HAPMIX (the highest performance of any method individually) to 99.770%. Combining calls from RFmix, LAMP-LD and HAPMIX did only slightly better than HAPMIX alone (99.755%) and the two groupings that omitted the calls of HAPMIX and LAMP-LD fared worse than HAPMIX. We also combined calls from all four of the methods in the study, but in this case only 98.712% of sites had a majority call while at the others there was a tie and no combined call could be made. This gave a considerable increase in performance to 99.915% which would be expected as the sites where a tie in the calls occurred are typically those at which is it harder to infer the ancestral population.

In the MXL3 simulations, where there are three possible ancestries at each locus, at most 99.967% of sites could be called by majority vote between three methods at a time. At the others the collective call was ambiguous with all three of the methods giving a different ancestry estimate. When combining calls across all four methods, 92.989% of sites could be called without ties occurring at which the combined calls had an accuracy of 98.964%. This level of accuracy was a considerable improvement to that achieved by any method individually or by any combination of the calls of three methods, however it leaves the ancestry of nearly 7% of sites undetermined.

4.5 Summary

Through these simulations we have compared the performance of four methods of local ancestry estimation - RFmix, LAMP-LD, MULTIMIX and HAPMIX - in both two and three-way admixture scenarios chosen to be representative of real admixed individuals. The set-up of these studies involved phased ancestral panels composed of haplotypes from the same population as those used in simulating the admixed individuals, and the phase of the samples was also known making it a favourable scenario for judging method performance. Over all, HAPMIX made the highest number of correct ancestry calls in both the ASW2 and the MXL3 groups, followed closely by LAMP-LD and MULTIMIX. The RFmix results were not as accurate as those of the other three methods, and it stood out as being considerably poorer than the others when various features of the estimates were studied.

In the ASW2 simulations LAMP-LD was the best at calling very narrow chunks, of less than 1000 SNPs in length, while in the MXL3 study HAPMIX was the best, doing slightly better than LAMP-LD. The LAMP-LD method, however, was also more susceptible to inferring spurious short chunks than MULTIMIX or HAPMIX were, but this error was the most common in the RFmix ancestry calls. The exact location of boundaries between regions of different ancestry were most accurately inferred by HAPMIX in both of the simulated groups, followed by MULTIMIX in the ASW2 study and by LAMP-LD in MXL3 tests. The RFmix method struggled the most in determining where the change occurred, and made two to three times more incorrect calls at boundaries than the other methods did.

We also found that combining calls via a majority vote across the estimates of LAMP, MULTIMIX and HAPMIX improved performance in the ASW2 simulations by 0.015% over that of HAPMIX alone. In the MXL3 simulations, the combined calls of these three methods also improved performance over that achieved by any method in-

dividually, however at a very small proportion (0.033%) of sites a consensus call could not be made in this manner.

These comparisons show that MULTIMIX is competitive with HAPMIX and LAMP-LD, in both two-way and multi-way admixture situations. They demonstrate that our method is a useful and flexible tool, based upon a novel model, that can deliver accurate ancestry estimates of the same caliber as the leading methods in the field.

Chapter 5

Analysis of Real Samples

The simulation studies of Chapters 3 and 4 show that our MULTIMIX method delivers accurate ancestry calls in range of situations, where admixture may arise from the mixing of more than two populations, and that its performance is in line with that of the leading methods in this field. Confirmation that the model has these qualities is important as there is an increasingly large number of Latino and Hispanic populations being involved in genome-wide studies. Owing to their colonial history, these groups typically display a range of inheritance of ancestry from three continents - Europe, Africa and the Americas - making them a more challenging cohort to study than, say, African Americans. In this chapter we apply the MULTIMIX model to genomes of African American, Mexican, Puerto Rican and Colombian individuals collected as part of the HapMap3 and 1000 Genomes projects. We use it to estimate the global ancestry proportions of each individual and to illustrate the switching in local ancestry across an admixed genome.

5.1 Mexican Individuals of HapMap3

The data of HapMap Phase 3 contains samples from 52 Mexicans (MEX) living in Los Angeles, California who have identified themselves as having at least three of their four grandparents born in Mexico. The CEU and YRI samples, also of HapMap, were used as the putative ancestral populations, selected to be representative of the European and West African ancestry of Mexicans. This mixed ancestry is a consequence of the history of Spanish settlers in the early sixteenth century and their introduction of slaves to the region (Cavalli-Sforza et al., 1994; Salzano and Bortolini, 2002).

In the analysis of these Mexican individuals, we used the HGDP (Cann et al., 2002) data set which contains samples from five Native American populations: 14 haplotypes from the Piapoco and Curripaco (collectively) of Colombia, 42 Maya and 28 Pima of Mexico, 26 Karitiana and 16 Surui of Brazil. The SNP set at which MULTIMIX was applied consisted of 589,246 biallelic SNPs genome-wide that were common to both the HapMap and HGDP data sets. A PCA analysis (Fig.5.1) of each of the Native American samples along with those of the CEU, YRI and the MEX gives evidence in support of the ancestral populations of the Mexicans being European, African and Native American, with most individuals lying along a cline between the European and Native American samples.

We analysed the phased MEX samples with EM-MULTIMIX at all combinations of a window size of $n \in \{100, 150\}$ SNPS and $\lambda \in \{0.005, 0.05, 0.5, 1\}$. Each run took about 50 minutes to complete. The haplotypes of the five Native American populations were combined to form one amalgamated Native American panel. In order to determine how well the ancestral segments that were inferred in each run corresponded to the ancestral populations used in the analysis, we calculated the Fst between the inferred chunks and the corresponding panel. This was done by making local ancestry calls for every MEX individual, giving us the chromosomal chunks estimated to belong to

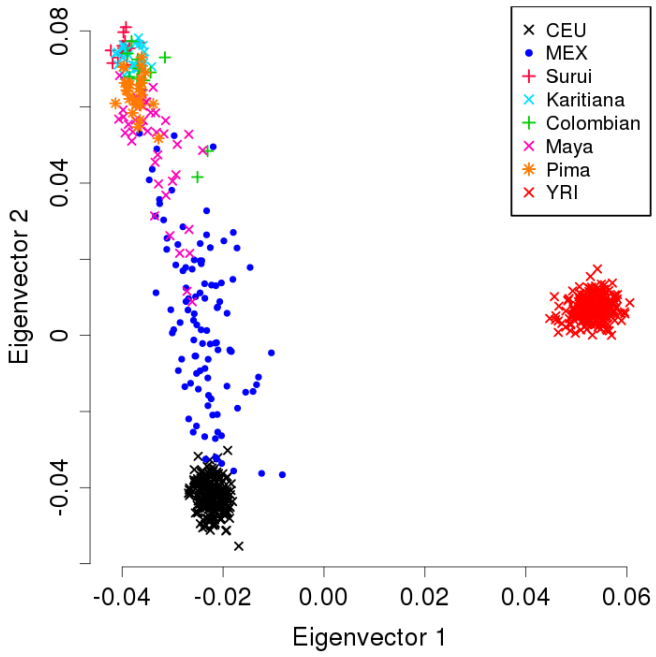

Figure 5.1: Principal components analysis of the CEU, YRI, MEX and all Native American populations (Surui, Karitiana, Colombian, Maya and Pima).

each ancestral population. The haplotypes inferred at these chunks were then compiled across the genome to give an inferred sample of haplotypes from each ancestral population as deduced from the MEX samples. This was done with the view that the run in which the parameter settings were optimal would lead to more accurate ancestry calls and therefore the lowest Fst between the inferred ancestral haplotypes and those in the panel. Only SNPs at which there were at least five chromosome copies amongst the MEX samples estimated to have ancestry from the relevant population at that site were used in the Fst calculation. In this analysis we found that the optimal parameters were $n = 150$ and $\lambda = 0.05$ and the Fst values for the European, African and Native American segments were 0.00466, 0.00509 and 0.0112 respectively. With these parameter settings, we also found that the complete-data log-likelihood of the MEX samples was largest over all runs. We repeated this analysis, but used only the Maya rather than all five of the Native American populations. In this case the Fst values

were slightly higher at 0.00465, 0.00509 and 0.0170 for the optimal parameter settings, suggesting that the amalgamated Native American panel was a more appropriate surrogate for the ancestral Native American haplotypes of the Mexican samples than the Maya alone.

The estimated global proportions for each of the 52 MEX individuals is displayed in order of increasing CEU ancestry in Fig.5.2. In this figure each vertical bar corresponds to one individual and is coloured in proportion to the inferred global ancestries of that person. It can be seen from this figure that we found the global ancestry proportions of the MEX individuals to be either predominantly European or Native American, with a very small contribution of West African ancestry. This observation is consistent with the analysis by Bryc et al. of the Mexican samples in the Population Reference Sample project (POPRES) in which FRAPPE was used to estimate global ancestry proportions at approximately 74,000 SNPs common to the Affymetrix 500K array and the Illumina 610-Quad panel (Bryc et al., 2010b). Furthermore, the average proportion of European, African and Native American ancestry and their standard deviations were estimated by MULTIMIX to be 0.518 (0.160), 0.0467 (0.0216) and 0.435 (0. 155) respectively, with the median values being 0.492, 0.0456 and 0.442. These results are in line with those found by Johnson et al. (2011) who analysed 46 of the Mexicans of HapMap using FRAPPE at a set of 482,906 SNPs. They estimated global ancestry proportions and reported very similar median values: 0.49, 0.05 and 0.45 respectively. It is reassuring to see that the application of our MULTIMIX method to real Mexican individuals gives similar results to previous analyses involving other Mexican samples.

5.2 1000 Genomes Admixed Samples

The 1000 Genomes Project provided samples collected from Americans of African ancestry in South West USA (ASW), Puerto Ricans living in Puerto Rico (PUR), Colom-

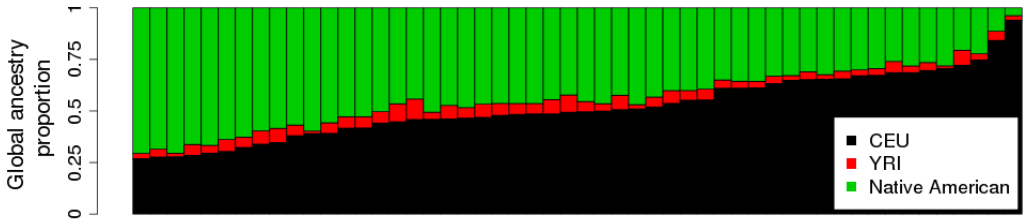

Figure 5.2: Global ancestry estimates for the 52 MEX individuals estimated by EM-MULTIMIX with ancestral populations CEU, YRI and the combined Native American populations. Each vertical bar displays the global ancestry proportions of one individual. A window size of $w = 150$ SNPs was used and $\lambda = 0.05$. First and third quantiles of the CEU, YRI and Native American proportions are $(0.418, 0.649)$, $(0.0326, 0.0563)$ and $(0.327, 0.528)$ respectively.

bians from Medellin Colombian (CLM) and individuals of Mexican ancestry living in Los Angeles (MXL). These samples, being part of the largest and most ambitious human genetic study to date, gave us the opportunity to apply our method to contemporary real data consisting of a considerable number of admixed individuals.

There were 65 African American individuals genotyped on the Omni platform at 2,177,885 SNPs across the genome. Their genotypes were phased using the statistical phasing algorithm SHAPE-IT (Delaneau et al., 2011). A total of 44 of these individuals were members of either trios or duos meaning that their phasing is expected to be highly accurate and reliable. For these individuals, we applied EM-MULTIMIX to their phased data to call the local ancestry of each haplotype separately. However, for the other 21 individuals who had no relatives amongst the ASW sample, their phasing is expected to be more susceptible to phase switch errors and so they were instead analysed as unphased genotype samples using the MCMC-MULTIMIX approach. The unrelateds were analysed in this way to avoid errors in haplotype estimation confounding the inference of local ancestry since it was noticed that in these samples there was far more ancestral switching as compared to the trio or duo-phased samples when analysed as phased haplotypes. Furthermore, the switches along pairs of chromosomes were in fact highly correlated. In analysing samples that may not be

very accurately phased it is a particularly useful feature of our MULTIMIX model that it can handle both haploid and diploid samples in this way.

In addition to the African Americans, we also analysed 64 Puerto Ricans (PUR), 66 Colombians (CLM) and 68 Mexican (MXL) individuals who were typed on the Affymetrix platform at 614,406 SNPs genome-wide. Most of these were collected as trios, with the exception of one CLM sample and 10 MXL samples. Their genotypes were also phased using SHAPE-IT and for those that were trio-phased, their phased haplotypic data were analysed via EM-MULTIMIX while the other 11 samples that were unrelateds were analysed as unphased diploid genotypes by MCMC-MULTIMIX, as was done in the analysis of the ASW. The ancestral panels that were used consisted of 762 European (EUR), 370 African (AFR) and 86 Native American (NATAM) haplotypes typed and phased at the same SNP set as the admixed samples. The Native American panel was composed of samples collected by Mao et al. which consisted of Maya individuals of the Yucatan Peninsula, Mexico; Nahau individuals from the state of Guerrero, Mexico; Aymara and Quechua individuals from La Paz; and Quechua individuals from Cerro de Pasco, Peru (Mao et al., 2007).

The inferred global ancestry proportions of the MXL, CLM and PUR individuals are displayed in Fig.5.3. We can see from the MXL plot that, as with the Mexican samples of HapMap that we analysed, the majority of the ancestry of the Mexican individuals is either European or Native American with only a very small African component. The average ancestral proportions for each admixed population are listed in Table 5.1. The PUR individuals show the highest average proportion of African ancestry, being nearly three times that found in the Mexicans, followed by the CLM which show on average twice as much African ancestry.

We can see from Fig.5.3 that there is considerable within-population variation in the ancestral proportions of the individuals of these groups, so while we cannot show what the ancestry of a "typical" Puerto Rican may look like, for instance, in Fig.5.4 we

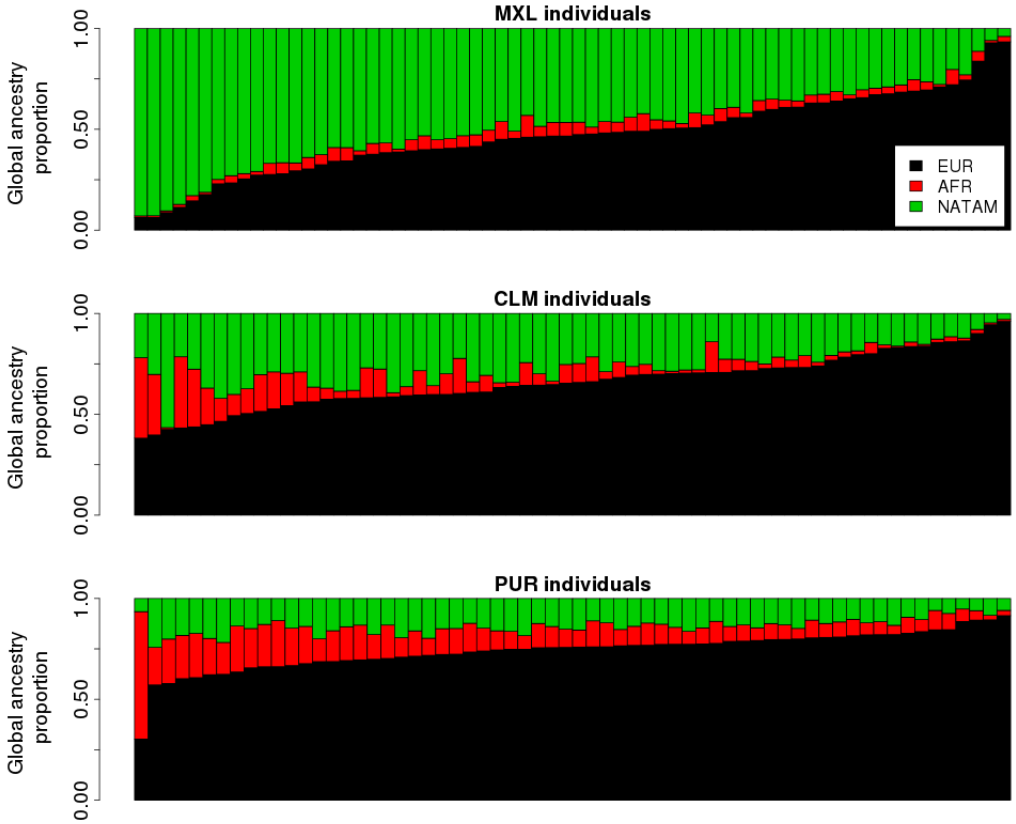

Figure 5.3: Global ancestry estimates for the PUR, CLM and MXL individuals of the 1000 Genomes Project.

Admixed samples	Estimated proportion of EUR ancestry (±s.d.)	Estimated proportion of AFR ancestry (±s.d.)	Estimated proportion of NATAM ancestry (±s.d.)
MEX of HapMap	0.518 (±0.160)	0.0467 (±0.0216)	0.435 (±0.155)
MXL of 1000 Genomes	0.470 (±0.192)	0.0440 (±0.0211)	0.486 (±0.196)
CLM of 1000 Genomes	0.659 (±0.133)	0.0797 (±0.0837)	0.261 (±0.0953)
PUR of 1000 Genomes	0.741 (±0.0950)	0.121 (±0.0837)	0.138 (±0.0389)

Table 5.1: The estimated global ancestry proportions of the MEX, MXL, CLM and PUR individuals as inferred by MULTIMIX.

give one example of an individual who exhibits a notable proportion of each of the three ancestries. This chromosome painting illustrates how the local ancestry inferred by our method varies along each pair of copies of the autosomes. The global ancestry proportions of EUR, AFR and Native American ancestry for this individual were found to be $(0.662, 0.228, 0.110)$ respectively.

Having estimated the local ancestry of the CLM, PUR and MXL samples, we asked how similar the European and African inferred tracts were to various European and African populations. The 1000 Genomes data set includes genotyped individuals from five European populations: Utah residents of north western European ancestry (CEU), Toscan individuals (TSI), Finnish individuals (FIN), British individuals from England and Scotland (GBR) and Iberian individuals in Spain (IBS). It also contains samples from three African populations: Yoruba of Ibadan in Nigeria (YRI), Luhya of Webuye in Kenya (LWK) and Maasai of Kinyawa in Kenya (MKK). These samples were genotyped on the Illumina platform providing a set of 239600 SNPs in common with those of the Affy platform on which the admixed samples were typed. Given the ancestry calls made by our MULTIMIX method, we calculated the Fst between the inferred tracts and each population of the same continental origin.

Table 5.2 shows the values of Fst found between the inferred European/African tracts of the three admixed populations and the corresponding 1000 Genomes populations. We found that the Spanish (IBS) samples were more closely related to the inferred European chunks of the CLM, PUR and MXL individuals than any of the other populations included in this analysis, and that the African tracts were the least diverged from the YRI than either of the Kenya populations. This is consistent with the history of the colonization of Colombia, Puerto Rico and Mexico primarly by the Spanish and their introduction of West African slaves to these regions (Salzano and Bortolini, 2002).

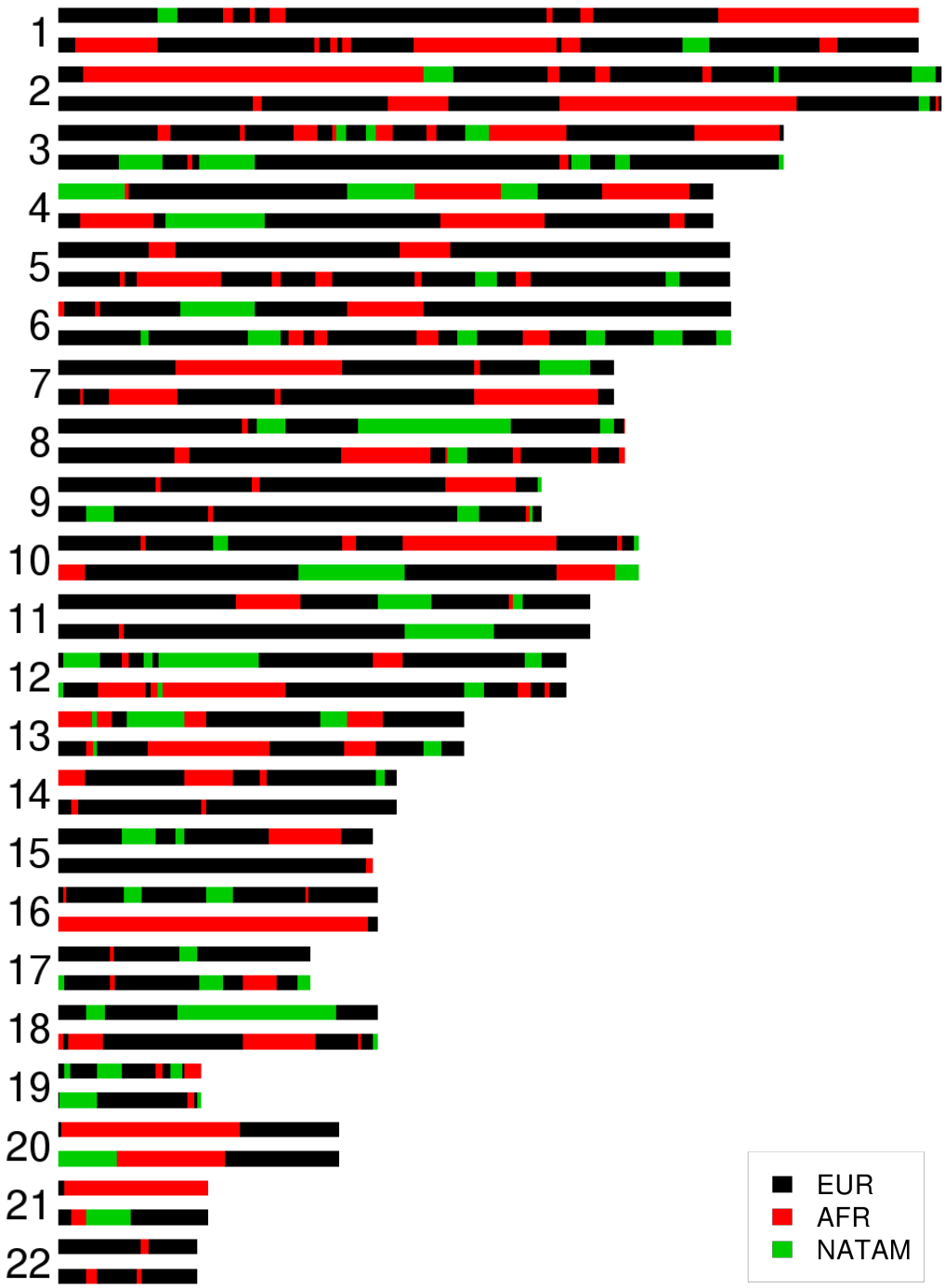

Figure 5.4: The chromosome painting of a single PUR individual inferred by EM-MULTIMIX. Both copies of all 22 autosomes are displayed, each pair labeled with their chromosome number.

Inferred ancestry	Admixed samples	IBS	TSI	CEU	GBR	FIN	YRI	LWK	MKK
EUR	CLM	0.00464	0.00512	0.00549	0.00569	0.00880			
EUR	PUR	0.00466	0.00504	0.00557	0.00575	0.00901			
EUR	MXL	0.00468	0.00495	0.00531	0.00545	0.00785			
AFR	CLM						0.00556	0.00598	0.0187
AFR	PUR						0.00640	0.00690	0.0203
AFR	MXL						0.00584	0.00615	0.0187

Table 5.2: The value of Fst found between the inferred European/African tracts of the admixed samples and several populations of corresponding continental ancestry.

5.3 Discussion

The purpose of this thesis is to put forward a new method, MULTIMIX, to estimate genome-wide local ancestry in admixed individuals from linked biallelic SNP data. The ability to accurately infer local ancestry is key to accounting for hidden population structure in disease association analyses, and to leveraging the variation in ancestry of admixed individuals who are cases for a particular phenotype within admixture mapping studies. It has also found use in the estimation of recombination maps, the search for signals of recent selection, and in studying population demographic history.

The importance of ancestry estimation is widely recognised and so there is already a series of proposed methods to tackle this problem. The more naive approaches such as the original STRUCTURE model, FRAPPE and ADMIXTURE apply to only a sparse set of unlinked markers because they do not model admixture LD - the pattern of ancestry switching that is characteristic of admixed genomes. A step above these models are approaches such as the STRUCTURE linkage model and ANCESTRYMAP which do model admixture LD but do not account for background LD between loci who alleles descend from the same ancestral population and are linked. SABER and SWITCH-MHMM are based upon first-order HMMs that go some way to accounting for background LD, while HAPMIX and LAMP-LD are the most sophisticated in this sense as

they model both admixture LD and background LD.

MULTIMIX makes use of dense genome-wide SNP data from the admixed samples being analysed and from panels of individuals representative of the candidate ancestral populations. It is a window-based approach in which the model of the probability of a haplotype given its population of ancestry follows an MVN distribution which effectively captures the LD structure of the haplotype within that window. The information at these contiguous windows is then combined via a hidden Markov model in which the probability of the ancestry switching from one window to another is a function of the age of the admixture and the global ancestry proportions. We include a post-processing step to resolve the precise location of the changes in ancestry within a window by considering all possible locations of the switch within a search region. By selecting the position which maximizes the likelihood of the admixed haplotype given the ancestry assignment, this final step allows us to refine the estimates on the ancestral population at SNPs located near these switches.

We described three methods to fit the model : an MCMC sampling scheme, an EM algorithm, and its variation the CEM algorithm. The performance of each method, measured by the percentage of correct ancestry calls, was tested via simulation studies involving both two-way and three-way admixture with samples simulated to exhibit a range of ancestry proportions and age of admixture. In a simulation of recently admixed CEU-YRI samples, we found the performance of our MULTIMIX method to be comparable to that of HAPMIX achieving an accuracy 99.724% correct calls, only 0.146% lower than HAPMIX. CEM and EM-MULTIMIX was found to be over twice as computationally fast as HAPMIX, analysing 400 copies of chromosome 1 in just over an hour.

Both the MULTIMIX and HAPMIX models are applicable to either phased or unphased admixed samples, but the MULTIMIX model is more flexible in that, unlike HAPMIX, it does not require that the ancestral panels contain phased haplotypes. Be-

ing able to use MULTIMIX in this range of settings is particularly useful if the phasing of the sample being analysed or the ancestral panels is unreliable. Phase switch errors in the sample can lead to spurious switches in the inferred local ancestry, a problem that can be avoided by instead applying the method to the unphased genotype data of the sample to estimate diploid ancestry. From our experience of applying MULTIMIX to the real admixed samples described in this chapter, we would advise against analysing samples as phased if their phasing was not conducted as part of either duos or trios. Such samples may have considerably higher switch error rates as a result of less accurate phasing that seem to cause spurious switches in local ancestry estimates.

In a broader comparison of methods, we examined how well MULTIMIX, HAPMIX, LAMP-LD and RFmix were able to call ancestry in two simulation studies: one involving European and African admixture, and the other of European, African and Native American admixture. The results of these comparisons show that HAPMIX made the most correct calls in both sets of simulations achieving an accuracy of 99.754% for the two-way and 97.561% for the three-way samples. MULTIMIX and LAMP-LD were very close behind making only approximately 0.095% and 0.096% more incorrect calls respectively in the two-way analysis and 0.642% and 0.774% more incorrect calls in the three-way analysis. The results of these simulations suggest that the RFmix method is not of the same caliber as the other three that were tested, as it achieved only 98.955% and 94.626% correct calls. The local ancestry calls of MULTIMIX, HAPMIX and LAMP-LD were combined by a majority vote at each locus and these combined calls outperformed those of HAPMIX by 0.016% in the two-way setting. In the three-way simulations, this approach improved performance to 97.683% at the 99.967% of sites at which combined call was not a tie between the possible ancestries. This suggests that in studies of real data it would be advantageous to make use of the choice of accurate ancestry estimation methods available to combine calls in this way rather than employing a single method alone.

With the performance of our method validated through a range of simulations, we then applied it to the analysis of data from real admixed individuals of the HapMap and 1000 Genomes Project. These admixed samples included Mexican, Colombian and Puerto Rican individuals whose local ancestries were estimated by MULTIMIX and gave average global proportions that are in line with those previously reported in other studies. These results illustrated the within-population variation in ancestry of each of these groups, and showed that the Mexican individuals possess a lower extent of African ancestry than the Colombians and Puerto Ricans as would be expected from their differing colonial histories.

As with any statistical model, assumptions have been made to make the model tractable and computationally feasible. One assumption of MULTIMIX is that the panels of ancestral haplotypes are representative of the true ancestors of the admixed samples. The effect of violations of this assumption were assessed in the two-way CEU-YRI simulation study of Section 3.2.5. It was found that using a panel of haplotypes from Tuscan instead of the CEU had very little effect leading to a drop in performance of only 0.111%. Using a panel of Kenyan haplotypes in place of the YRI panel lead to a more considerable drop of 0.463%, however in this setting MULTIMIX outperformed HAPMIX. When both the Tuscan and Kenyan panels were used simultaneously, MULTIMIX made more correct calls than when only the Kenyan panel was substituted. It is interesting to note that the disparity between the performance of HAPMIX and MULTIMIX was considerably narrowed from 0.621% to 0.185% when both of the ancestral panels were inaccurate suggesting that our model is more resilient to misspecified or inaccurate ancestral panels. This might be due to the fact that HAPMIX explicitly models the admixed individual to be a composite of the panel haplotypes, while MULTIMIX uses only summary statistics (the population allele frequencies and their covariances) to describe the relative probability of an admixed haplotype given it has one ancestry as opposed to another. In the MULTIMIX model, it is the relative values of

the density of the admixed haplotype given the different ancestries that is important, rather than their actual values.

Another assumption of the MULTIMIX model is that the rate of switching between ancestry states per unit genetic distance the same across the whole genome. This may not be the case if there has been more than one admixture event in the population's history. We can imagine a scenario in which two populations mix at a particular point in time and admixture continues for several generations before a third population joins at another admixture event. In this case we would expect the genomes of the three-way admixed individuals to possess regions in which there are longer stretches of ancestry from the third population, and other areas along which a much higher rate of switching is observed between ancestries of the two original populations as more recombination events have occurred here over a greater number of generations. How considerable an effect this violation would have on the performance of our method has not been examined here.

5.4 Conclusion

In this thesis we have presented a novel method named MULTIMIX that uses bial-lelic SNP data to infer the local genetic ancestry in admixed individuals. At the time of completing this thesis, a journal paper describing the method has been accepted for publication in Genetic Epidemiology. While many ancestry estimation methods have been developed over the past ten years, MULTIMIX is the most flexible in that it can handle multi-way admixture while remaining computationally efficient even when analysing the whole genome at once, and furthermore may be applied to any scenario where the admixed samples and/or panels are phased or unphased. Furthermore it does not require that any parameters of the model such as the number of generations since admixture or the global admixture proportions are known as all model param-

eters are estimated. It is interesting to discover that using a purely statistical model, the MVN model on the distribution of haplotypes given the population of ancestry, can be competitive with approaches such as HAPMIX and LAMP-LD which explicitly model haplotypes via the Li and Stephens model using haplotypic phase information. With increasingly ambitious efforts such as the 1000 Genomes Project delivering genetic data on a growing number of admixed populations, there is a need for a method such as ours that can deliver fast and accurate ancestry inference in a wide extent of scenarios.

Bibliography

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

David H. Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

D.M. Altshuler, R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F. Yu, P.E. Bonnen, P.I. de Bakker, P. Deloukas, S.B. Gabriel, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

Y. Baran, B. Pasaniuc, S. Sankararaman, D.G. Torgerson, C. Gignoux, C. Eng, W. Rodriguez-Cintron, R. Chapela, J.G. Ford, P.C. Avila, et al. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, 28(10).

G. Barbujani and RR Sokal. Genetic population structure of Italy. II. Physical and cultural barriers to gene flow. *American journal of human genetics*, 48(2):398–411, 1991.

J.A. Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to parameter Estimation for Gaussian Mixture and Hidden Markov Models. 1998.

M.J. Bishop and C. Cannings. *Handbook of statistical genetics*, volume 1. Wiley-Interscience, 2007.

Cathryn Bock, Ann Schwartz, Julie Ruterbusch, Albert Levin, Christine Neslund-Dudas, Susan Land, Angela Wenzlaff, David Reich, Paul Mckeigue, Wei Chen, Elisa-

beth Heath, Isaac Powell, Rick Kittles, and Benjamin Rybicki. Results from a prostate cancer admixture mapping study in African-American men. *Human Genetics*, 126(5): 637–642, 2009.

S.R. Browning and B.L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5):1084–1097, 2007.

Katarzyna Bryc, Adam Auton, Matthew R. Nelson, Jorge R. Oksenberg, Stephen L. Hauser, Scott Williams, Alain Froment, Jean-Marie Bodo, Charles Wambebe, Sarah A. Tishkoff, and Carlos D. Bustamante. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*, 107(2):786–791, 2010a.

Katarzyna Bryc, Christopher Velez, Tatiana Karafet, Andres Moreno-Estrada, Andy Reynolds, Adam Auton, Michael Hammer, Carlos D. Bustamante, and Harry Ostrer. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences*, 107 (Supplement 2):8954–8961, 2010b.

H.M. Cann, C. de Toma, L. Cazes, M.F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, et al. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

R.L. Cann, M. Stoneking, and A.C. Wilson. Mitochondrial DNA and human evolution. *Nature*, 325:31–36, 1987.

L. Luca Cavalli-Sforza, P. Menozzi, and A. Piazza. *The History and Geography of Human Genes* . Princeton University Press, 1994.

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.

Ching-Yu Cheng, Linda W. H. Kao, Nick Patterson, Arti Tandon, Christopher A. Haiman, Tamara B. Harris, Chao Xing, Esther M. John, Christine B. Ambrosone, Frederick L. Brancati, Josef Coresh, Michael F. Press, Rulan S. Parekh, Michael J. Klag, Lucy A. Meoni, Wen-Chi Hsueh, Laura Fejerman, Ludmila Pawlikowska, Matthew L. Freedman, Lina H. Jandorf, Elisa V. Bandera, Gregory L. Ciupak, Michael A. Nalls, Ermeg L. Akylbekova, Eric S. Orwoll, Tennille S. Leak, Iva Miljkovic, Rongling Li, Giske Ursin, Leslie Bernstein, Kristin Ardlie, Herman A. Taylor, Eric Boerwinckle, Joseph M. Zmuda, Brian E. Henderson, James G. Wilson, and David Reich. Admixture Mapping of 15,280 African Americans Identifies Obesity Susceptibility Loci on Chromosomes 5 and X. *PLoS Genet*, 5(5):e1000490, 2009.

Ching-Yu Y. Cheng, David Reich, Josef Coresh, Eric Boerwinkle, Nick Patterson, Man Li, Kari E. North, Arti Tandon, Joan E. Bailey-Wilson, James G. Wilson, and Linda H. Kao. Admixture mapping of obesity-related traits in African Americans: the atherosclerosis risk in communities (ARIC) Study. *Obesity (Silver Spring, Md.)*, 18 (3):563–572, 2010.

O. Delaneau, J. Marchini, and J.F. Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179–181, 2011.

Steven C. Elbein, Swapan K. Das, Michael D. Hallman, Craig L. Hanis, and Sandra J. Hasstedt. Genome-Wide Linkage and Admixture Mapping of Type 2 Diabetes in African American Families from the American Diabetes Association GENNID Cohort. *Diabetes*, (1), 2009.

Daniel Falush, Matthew Stephens, and Jonathan K. Pritchard. Inference of Popula-

tion Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*, 164(4):1567–1587, 2003.

Laura Fejerman, Christopher A. Haiman, David Reich, Arti Tandon, Rahul C. Deo, Esther M. John, Sue A. Ingles, Christine B. Ambrosone, Dana H. Bovbjerg, Lina H. Jandorf, Warren Davis, Gregory Ciupak, Alice S. Whittemore, Michael F. Press, Giske Ursin, Leslie Bernstein, Scott Huntsman, Brian E. Henderson, Elad Ziv, and Matthew L. Freedman. An Admixture Scan in 1,484 African American Women with Breast Cancer. *Cancer Epidemiology Biomarkers & Prevention*, 18(11):3110–3117, 2009.

Matthew L. Freedman, Christopher A. Haiman, Nick Patterson, Gavin J. Mcdonald, Arti Tandon, Alicja Waliszewska, Kathryn Penney, Robert G. Steen, Kristin Ardlie, Esther M. John, Ingrid Oakley-Girvan, Alice S. Whittemore, Kathleen A. Cooney, Sue A. Ingles, David Altshuler, Brian E. Henderson, and David Reich. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *PNAS*, 103(38):14068–14073, 2006.

M. Hammer, Y Chromosome Consortium, et al. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Research*, 12(2):339–348, 2002.

A.G. Hinch, A. Tandon, N. Patterson, Y. Song, N. Rohland, C.D. Palmer, G.K. Chen, K. Wang, S.G. Buxbaum, E.L. Akylbekova, et al. The landscape of recombination in African Americans. *Nature*, 476(7359):170–175, 2011.

L. Hirszfeld and H. Hirszfeld. Essai d'application des méthodes sérologiques au problème des race. *Anthropologie*, (29):505–537, 1919.

B.N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, 2009.

N.A. Johnson, M.A. Coram, M.D. Shriver, I. Romieu, G.S. Barsh, S.J. London, and H. Tang. Ancestral Components of Admixed Genomes in a Mexican Cohort. *PLoS Genetics*, 7(12):e1002410, 2011.

Linda H. Kao, Michael J. Klag, Lucy A. Meoni, David Reich, Yvette Berthier-Schaad, Man Li, Josef Coresh, Nick Patterson, Arti Tandon, Neil R. Powe, Nancy E. Fink, John H. Sadler, Matthew R. Weir, Hanna E. Abboud, Sharon G. Adler, Jasmin Divers, Sudha K. Iyengar, Barry I. Freedman, Paul L. Kimmel, William C. Knowler, Orly F. Kohn, Kristopher Kramp, David J. Leehey, Susanne B. Nicholas, Madeleine V. Pahl, Jeffrey R. Schelling, John R. Sedor, Denyse Thornley-Brown, Cheryl A. Winkler, Michael W. Smith, Rulan S. Parekh, and Family Investigation of Nephropathy and Diabetes Research Group. MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nature genetics*, 40(10):1185–1192, 2008.

T.M. Karafet, F.L. Mendez, M.B. Meilerman, P.A. Underhill, S.L. Zegura, and M.F. Hammer. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research*, 18(5):830–838, 2008.

A. Kong, G. Thorleifsson, D.F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G.B. Walters, A. Jonasdottir, A. Gylfason, K.T. Kristinsson, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467 (7319):1099–1103, 2010.

J. De Leeuw. Block relaxation algorithms in statistics. *Information systems and data analysis*, pages 308–324, 1994.

Na Li and Matthew Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165 (4):2213–2233, 2003.

Y. Li, C.J. Willer, J. Ding, P. Scheet, and G.R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34:816–834, 2010.

X. Mao, A.W. Bigham, R. Mei, G. Gutierrez, Weiss K.M., T.D. Brutsaert, F. Leon-Velarde, L.G. Moore, E. Vargas, P.M. McKeigue, M.D. Shriver, and E.J. Parra. A Genomewide Admixture Mapping Panel from Hispanic/Latino Populations. *American Journal of Human Genetics*, 80(6):1171–1178, 2007.

Jonathan Marchini, Lon R. Cardon, Michael S. Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5):512–517, 2004.

E. Mayr. *Systematics and the Origin of Species*. Columbia University Press, 1942.

I. McDougall, F.H. Brown, and J.G. Fleagle. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, 433(7027):733–736, 2005.

P. Menozzi, A. Piazza, and L. Cavalli-Sforza. Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–792, 1978.

S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, 2005.

S. Myers, C. Freeman, A. Auton, P. Donnelly, and G. McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature genetics*, 40(9):1124–1129, 2008.

M. Nelis, T. Esko, et al. Genetic Structure of Europeans: a view from the North-East. *Plos One*, 4(5):e5472, 2009.

M.R. Nelson, K. Bryc, K.S. King, A. Indap, A.R. Boyko, J. Novembre, L.P. Briley, Y. Maruyama, D.M. Waterworth, G. Waeber, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3):347–358, 2008.

J Novembre, T Johnson, K Bryc, Z Kutalik, A R Boyko, A Auton, A Indap, K S King, S Bergmann, M R Nelson, M Stephens, and C D Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

Bogdan Pasaniuc, Sriram Sankararaman, Gad Kimmel, and Eran Halperin. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):213–221, 2009.

N. Patterson, N. Hattangadi, B. Lane, K. Lohmueller, D. Hafler, J. Oksenberg, S. Hauser, M. Smith, S. Obrien, and D. Altshuler. Methods for High-Density Admixture Mapping of Disease Genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.

J.E. Pool and R. Nielsen. Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. *Genetics*, 181(2):711–719, 2009.

A. L. Price, N. Patterson, F. Yu, D. R. Cox, A. Waliszewska, G. J. McDonald, A. Tandon, C. Schirmer, J. Neubauer, G. Bedoya, C. Duque, A. Villegas, M. C. Bortolini, F. M. Salzano, C. Gallo, G. Mazzotti, M. Tello-Ruiz, L. Riba, C. A. Aguilar-Salinas, S. Canizales-Quinteros, M. Menjivar, W. Klitz, B. Henderson, C. A. Haiman, C. Winkler, T. Tusie-Luna, A. Ruiz-Linares, and D. Reich. A genomewide admixture map for Latino populations. *American journal of human genetics*, 80(6):1024–1036, 2007.

A.L. Price, M.E. Weale, N. Patterson, S.R. Myers, A.C. Need, K.V. Shianna, D. Ge, J.I. Rotter, E. Torres, K.D. Taylor, et al. Long-range LD can confound genome scans in admixed populations. *American journal of human genetics*, 83(1):132–135, 2008.

Alkes L. Price, Arti Tandon, Nick Patterson, Kathleen C. Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H. Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, 5(6):e1000519, 2009.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. De Bakker, M.J. Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

S. Ramachandran, O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman, and L.L. Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942, 2005.

D. Reich, K. Thangaraj, N. Patterson, A.L. Price, and L. Singh. Reconstructing Indian population history. *Nature*, 461(7263):489–494, 2009.

B. Ringelhann, M.K. Hathorn, P. Jilly, F. Grant, and G. Parniczky. A new look at the protection of hemoglobin AS and AC genotypes against plasmodium falciparum infection: a census tract approach. *American Journal of Human Genetics*, 28(3):270–279, 1976.

N.A. Rosenberg, L. Huang, E.M. Jewett, Z.A. Szpiech, I. Jankovic, and M. Boehnke. *Nature Reviews Genetics*, 11(5):356–366, 2010.

F.M. Salzano and M.C. Bortolini. *The evolution and genetics of Latin American populations*, volume 28. 2002.

Sriram Sankararaman, Gad Kimmel, Eran Halperin, and Michael I. Jordan. On the inference of ancestries in admixed populations. *Genome Research*, 18(4):668–675, 2008a.

Sriram Sankararaman, Srinath Sridhar, Gad Kimmel, and Eran Halperin. Estimating Local Ancestry in Admixed Populations. *The American Journal of Human Genetics*, 82 (2):290–303, 2008b.

M.W. Smith, N. Patterson, J.A. Lautenberger, A.L. Truelove, McDonald G.J., A. Waliszewska, B.D. Kessing, et al. A high-density admixture map for disease gene discovery in African Americans. *American Journal of Human Genetics*, 74(5):1001–1013, 2004.

A. Sundquist, E. Fratkin, C.B. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome research*, 18(4):676–682, 2008.

H. Tang, S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Cintron, E.G. Burchard, and N.J. Risch. Recent genetic selection in the ancestral admixture of Puerto Ricans. *The American Journal of Human Genetics*, 81(3):626–633, 2007.

Hua Tang, Jie Peng, Pei Wang, and Neil J. Risch. Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28:289–301, 2005.

Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing genetic ancestry blocks in admixed individuals. *American journal of human genetics*, 79(1):1–12, 2006.

J.D. Wall, R. Jiang, C. Gignoux, G.K. Chen, C. Eng, S. Huntsman, and P. Marjoram. Genetic variation in Native Americans, inferred from Latino SNP and resequencing data. *Molecular Biology and Evolution*, 28(8):2231–2237, 2011.

D. Wegmann, D.E. Kessner, K.R. Veeramah, R.A. Mathias, D.L. Nicolae, L.R. Yanek, Y.V. Sun, D.G. Torgerson, N. Rafaels, T. Mosley, et al. Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics*, 43(9):847–853, 2011.

Bruce Weir. *Genetic Data Analysis 2: Methods for Discrete Population Genetic Data*. Sinauer Associates Inc, 1996.

X. Wen and M. Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, 4(3):1158–1182, 2010.

S. Wright. The theoretical variance within and among subdivisions of a population that is in a steady state. *Genetics*, 37(3):312–321, 1952.

X. Zhu and R. S. Cooper. Admixture mapping provides evidence of association of the VNN1 gene with hypertension. *PLoS ONE*, 2(11):e1244, 2007.

X. Zhu, A. Luke, R. S. Cooper, T. Quertermous, C. Hanis, T. Mosley, C. C. Gu, H. Tang, D. C. Rao, N. Risch, and A. Weder. Admixture mapping for hypertension loci with genome-scan markers. *Nat Genet*, 37(2):177–181, 2005.